

Encyclopedia of Information Science and Technology, Fifth Edition

Mehdi Khosrow-Pour D.B.A.
Information Resources Management Association, USA



A volume in the Advances in Information Quality
and Management (AIQM) Book Series

Published in the United States of America by

IGI Global
Engineering Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: <http://www.igi-global.com>

Copyright © 2021 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher. Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Names: Khosrow-Pour, Mehdi, 1951- editor.

Title: Encyclopedia of information science and technology / Mehdi

Khosrow-Pour, D.B.A., editor.

Description: Fifth edition. | Hershey, PA : Engineering Science Reference, 2020. | Includes bibliographical references and index. | Summary: "This book presents multidisciplinary research and expert insights into new methods and processes for understanding modern technological tools and their applications as well as emerging theories and ethical controversies surrounding the field of information science"-- Provided by publisher.

Identifiers: LCCN 2020000232 (print) | LCCN 2020000233 (ebook) | ISBN 9781799834793 (hardcover) | ISBN 9781799834809 (ebook)

Subjects: LCSH: Information science--Encyclopedias. | Information technology--Encyclopedias.

Classification: LCC Z1006 .E566 2020 (print) | LCC Z1006 (ebook) | DDC 020.3--dc23

LC record available at <https://lcn.loc.gov/2020000232>

LC ebook record available at <https://lcn.loc.gov/2020000233>

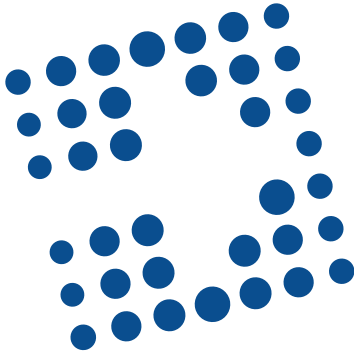
This book is published in the IGI Global book series Advances in Information Quality and Management (AIQM) (ISSN: 2331-7701; eISSN: 2331-771X)

British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

For electronic access to this publication, please contact: eresources@igi-global.com.



Advances in Information Quality and Management (AIQM) Book Series

Siddhartha Bhattacharyya
RCC Institute of Information Technology, India

ISSN:2331-7701
EISSN:2331-771X

MISSION

Acquiring and managing quality information is essential to an organization's success and profitability. Innovation in information technology provides managers, researchers, and practitioners with the tools and techniques needed to create and adapt new policies, strategies, and solutions for information management.

The **Advances in Information Quality and Management (AIQM) Book Series** provides emerging research principals in knowledge society for the advancement of future technological development. This series aims to increase available research publications and emphasize the global response within the discipline and allow for audiences to benefit from the comprehensive collection of this knowledge.

COVERAGE

- Human and Societal Issue
- Knowledge Management
- Decision Support and Group Decision Support Systems
- Electronic Commerce Technologies
- Supply Chain Management
- Business Process Management and Modeling
- Application of IT to Operation
- Emerging Technologies Management
- IT Innovation and Diffusion
- Web Services and Technologies

IGI Global is currently accepting manuscripts for publication within this series. To submit a proposal for a volume in this series, please contact our Acquisition Editors at Acquisitions@igi-global.com or visit: <http://www.igi-global.com/publish/>.

The Advances in Information Quality and Management (AIQM) Book Series (ISSN 2331-7701) is published by IGI Global, 701 E. Chocolate Avenue, Hershey, PA 17033-1240, USA, www.igi-global.com. This series is composed of titles available for purchase individually; each title is edited to be contextually exclusive from any other title within the series. For pricing and ordering information please visit <http://www.igi-global.com/book-series/advances-information-quality-management/73809>. Postmaster: Send all address changes to above address. Copyright © 2021 IGI Global. All rights, including translation in other languages reserved by the publisher. No part of this series may be reproduced or used in any form or by any means – graphics, electronic, or mechanical, including photocopying, recording, taping, or information and retrieval systems – without written permission from the publisher, except for non commercial, educational use, including classroom teaching purposes. The views expressed in this series are those of the authors, but not necessarily of IGI Global.

Titles in this Series

For a list of additional titles in this series, please visit: www.igi-global.com/book-series

Handbook of Research on Managing Information Systems in Developing Economies

Richard Boateng (University of Ghana, Ghana)

Business Science Reference • ©2020 • 695pp • H/C (ISBN: 9781799826101) • US \$265.00

Legal and Ethical Issues in Information Management

Peter Mazebe II Mothataesi Sebina (University of Botswana, Botswana) Nathan Mwakoshi Mnjama (University of Botswana, Botswana) Kgomotso Hildegard Moahi (University of Botswana, Botswana) Tachilisa Badala Balule (University of Botswana, Botswana) and Balulwami Grand (University of Botswana, Botswana)

Information Science Reference • ©2020 • 320pp • H/C (ISBN: 9781799824220) • US \$195.00

Metaheuristic Approaches to Portfolio Optimization

Jhuma Ray (RCC Institute of Information Technology, India) Anirban Mukherjee (RCC Institute of Information Technology, India) Sadhan Kumar Dey (RCC Institute of Information Technology, India) and Goran Klepac (Algebra University College, Croatia)

Business Science Reference • ©2019 • 263pp • H/C (ISBN: 9781522581031) • US \$205.00

Strategic Management of Business-Critical Information Assets

Denise A.D. Bedford (Georgetown University, USA)

Business Science Reference • ©2019 • 308pp • H/C (ISBN: 9781522584100) • US \$205.00

Systems Research for Real-World Challenges

Frank Stowell (University of Portsmouth, UK)

Information Science Reference • ©2018 • 398pp • H/C (ISBN: 9781522559962) • US \$195.00

Ensuring Research Integrity and the Ethical Management of Data

Cees Th. Smit Sibinga (University of Groningen, The Netherlands)

Information Science Reference • ©2018 • 303pp • H/C (ISBN: 9781522527305) • US \$275.00

Handbook of Research on Global Enterprise Operations and Opportunities

Mehdi Khosrow-Pour, D.B.A. (Information Resources Management Association, USA)

Business Science Reference • ©2017 • 375pp • H/C (ISBN: 9781522522454) • US \$245.00

Examining Information Retrieval and Image Processing Paradigms in Multidisciplinary Contexts

Joan Lu (University of Huddersfield, UK) and Qiang Xu (University of Huddersfield, UK)

Information Science Reference • ©2017 • 425pp • H/C (ISBN: 9781522518846) • US \$245.00



701 East Chocolate Avenue, Hershey, PA 17033, USA

Tel: 717-533-8845 x100 • Fax: 717-533-8661

E-Mail: cust@igi-global.com • www.igi-global.com

This book is dedicated to the memory of my late father for the love and care that he always displayed for his family, and for teaching me the importance of humanitarianism. Also, to my beautiful wife, Olga, and our handsome sons, Darius and Cyrus, for filling my life with so much love, joy, and happiness.

Editorial Advisory Board

Gilad Ariely, *Lauder School of Government, Diplomacy, and Strategy, IDC Herzliya, Israel*
Md. Zahir Uddin Arif, *Jagannath University, India*
Sergey Balandin, *FRUCT Oy, Finland*
Carlos Ballesteros, *Universidad Pontificia Comillas, Spain*
William Bart, *University of Minnesota, USA*
Siddhartha Bhattacharyya, *RCC Institute of Information Technology, Kolkata, India*
Michael A. Brown Sr., *Florida Institute of Technology, USA*
Isabell Carolina Camillo, *Goodman School of Business, Brock University, St. Catharines, Canada*
Zhiyuan Chen, *University of Maryland, Baltimore County, USA*
Linda Daniela, *University of Latvia, Latvia*
Shirley M. Fedorovich, *Embry-Riddle Aeronautical University, USA*
Francisco José García-Peñalvo, *University of Salamanca, Spain*
Babu George, *Fort Hays State University, USA*
Vicente González-Prida, *University of Seville, Spain & UNED, Spain*
Maki K. Habib, *American University of Cairo, Egypt*
Shalin Hai-Jew, *Kansas State University, USA*
Telmo Antonio Henriques, *ISCTE, University Institute of Lisbon, Portugal*
Wen-Chen Hu, *University of North Dakota, USA*
Dariusz Jacek Jakobczak, *Koszalin University of Technology, Poland*
Kai Jakobs, *RWTH Aachen University, Germany*
George Leal Jamil, *Informações em Rede Consultoria e Treinamento, Brazil*
Murray E. Jennex, *San Diego State University, USA*
Hans Ruediger Kaufmann, *University of Applied Management Studies Mannheim, Germany & University of Nicosia, Cyprus*
Dimitry Korzun, *Petrozavodsk State University, Russia*
Rocci Luppigini, *University of Ottawa, Canada*
Kathryn MacCullum, *Eastern Institute of Technology, New Zealand*
Minna M. Mattila, *Embry-Riddle Aeronautical University, USA*
Elspeth McKay, *RMIT University, Australia*
Ionica Oncioiu, *Titu Maiorescu, Romania*
Patricia Ordonez de Pablos, *University of Oviedo, Spain*

Drew Polly, *University of North Carolina at Charlotte, USA*
Barre Vijaya Prasad, *All India Institute of Medical Sciences (AIIMS), India*
Hayo Reinders, *King Mongkut's University of Technology Thonburi, Thailand*
James Allen Rodger, *Indiana University of Pennsylvania, USA*
Joao M. F. Rodrigues, *University of Algarve, Portugal*
Manuel Pedro Rodríguez Bolívar, *University of Granada, Spain*
Aleksandra Rosiek-Kryszewska, *Nicolaus Copernicus University in Toruń, Poland & Collegium Medicum in Bydgoszcz, Poland*
Neil Rowe, *Naval Postgraduate School, USA*
Fjodor Ruzic, *HR Open, Croatia*
Ari Sigal, *Catawba Valley Community College, USA*
Emily Stacey, *Rose State College, USA*
Sang-Bing Tsai, *University of Electronic Science and Technology of China, China & Zhongshan Institute, China & Civil Aviation University of China, China & School of Business, Wuyi University, China*
Anna Ursyn, *University of Northern Colorado, USA*
Chelsea C. White, *Georgia Institute of Technology, USA*
Recep Yilmaz, *Ondokuz Mayıs Üniversitesi, Turkey*
Ali Hussain Saleh Zolait, *University of Bahrain, Bahrain*

List of Contributors

Abdul-Jeleel, Balikis Animasaun / <i>International Institute of Tropical Agriculture (IITA), Nigeria</i>	1699
Abusaada, Hisham / <i>Housing and Building National Research Center (HBRC), Egypt</i>	1196
Acosta, Olga / <i>Singularyta SpA, Chile</i>	856
Adedayo, Adekunbi Enoch / <i>Adeyemi Federal University of Education, Ondo, Nigeria</i>	1362
Afolabi, Asani M. / <i>Ladoke Akintola University, Nigeria</i>	1336
Agrawal, Nitesh Kumar / <i>Vellore Institute of Technology, India</i>	98
Aguilar, César / <i>Pontificia Universidad Católica de Chile, Chile</i>	856
Ahmed, Rukhsana / <i>University at Albany, SUNY, USA</i>	1839
Akinsumbo, Kolawole Sunday / <i>LucrativeTech, Nigeria</i>	1699
Aligwa, Rashida Mebude / <i>Tek Experts Limited, Nigeria</i>	1699
Almeida, Fernando / <i>Polytechnic Institute of Gaya, Portugal</i>	1513
Almunawar, Mohammad Nabil / <i>Universiti Brunei Darussalam, Brunei</i>	1683
Amoko, Justinah S. / <i>Adeyemi Federal University of Education, Ondo, Nigeria</i>	1362
Artese, Maria Teresa / <i>IMATI-CNR, Italy</i>	107
Artyukhov, Artem / <i>Sumy State University, Ukraine</i>	1289
Artyukhova, Nadiia / <i>Sumy State University, Ukraine</i>	1289
Asani, Afolabi M. / <i>Ladoke Akintola University, Nigeria</i>	1231
Averweg, Udo Richard / <i>eThekwini Municipality, South Africa</i>	1672
Babalola, Samuel Oyelami / <i>UAC Nigeria Plc, Nigeria</i>	1699
Babcock, Grace M. / <i>North Carolina State University, USA</i>	1096
Banik, Abira / <i>Tripura University, India</i>	1018
Barrera, María P. / <i>Universidad de Córdoba, Colombia</i>	54
Barrett, Diane / <i>Bloomsburg University of Pennsylvania, USA</i>	462
Ben Ouannes, Achref / <i>ARSELA, Tunisia</i>	1578
Bezboruah, Tulshi / <i>Gauhati University, India</i>	1037
Bhagat, Amit / <i>Maulana Azad National Institute of India, India</i>	559
Bhagwatikar, Govind / <i>Sany Wind Energy India Private Limited, India</i>	815
Boido, Cristina / <i>Politecnico di Torino, Italy</i>	1181
Bora, Abhijit / <i>Gauhati University, India</i>	1037
Borhani, Roula / <i>OMEDRH, Tunis, Tunisia</i>	1578
Bose, Indranil / <i>Indian Institute of Management, Kolkata, India</i>	1106
Bouchiha, Djelloul / <i>Center University of Naama, Algeria & EEDIS Lab. UDL-SBA, Algeria</i>	1214
Bratsas, Charalampos / <i>Aristotle University of Thessaloniki, Greece</i>	1140
Brown, Sheldondra J. / <i>Southeast Missouri State University, USA</i>	1096

C., Abinaya / <i>Thiagarajar College of Engineering, India</i>	473
Cahlik, Tomas / <i>Charles University, Prague, Czech Republic</i>	1451
Campos, Pedro / <i>INESC TEC, University of Porto, Portugal</i>	1827
Caro, Manuel F. / <i>Universidad de Córdoba, Colombia</i>	54
Casas-Rosal, Jose Carlos / <i>Universidad de Córdoba, Spain</i>	624
Chabukswar, Anuruddha R. / <i>School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India</i>	1797
Chaudhary, Pankaj / <i>Indiana University of Pennsylvania, USA</i>	1550
Chauhan, Nutan Kumari / <i>KIET Group of Institutions, India</i>	233
Chen, E. Jack / <i>BASF Corporation, USA</i>	164
Chhaya, Lipi / <i>Researcher, University of Petroleum and Energy Studies, India</i>	815
Choubey, Dilip Kumar / <i>Vellore Institute of Technology, Vellore, India</i>	980
Chowdhury, Mahfuzulhoq / <i>Institut National de la Recherche Scientifique (INRS), Canada & University of Quebec, Canada</i>	27
Cook, Bruce Lawrence / <i>Cook Communication, USA</i>	1461
Cordeiro, M. Natália D. S. / <i>LAQV-REQUIMTE, University of Porto, Portugal</i>	1827
Costa, Vera / <i>FEUP, University of Porto, Portugal</i>	637
Crossan, Adam / <i>Letterkenny Institute of Technology, Ireland</i>	1719
Curran, Kevin / <i>Ulster University, UK</i>	1719
Das, Arpita / <i>University of Calcutta, India</i>	271
Das, Kaustav / <i>University of Calcutta, India</i>	271
Dasso, Aristides / <i>Universidad Nacional de San Luis, Argentina</i>	749
Davico, Pia / <i>Politecnico di Torino, Italy</i>	1181
Delgado, José Carlos Martins / <i>Instituto Superior Técnico, Universidade de Lisboa, Portugal</i>	217
Dhar, Suparna / <i>NSHM Knowledge Campus, Kolkata, India</i>	1106
Dikovic, Ljubica / <i>College of Applied Sciences Uzice, Serbia</i>	1933
Dimou, Anastasios / <i>Independent Researcher, Athens, Greece</i>	1118
Donaj, Gregor / <i>University of Maribor, Slovenia</i>	73
Doorn, Jorge Horacio / <i>Universidad Nacional de La Matanza, Argentina & Universidad Nacional de Tres de Febrero, Argentina</i>	1165
Duarte, Dinah / <i>CBIOS, Universidade Lusófona, Portugal</i>	1815
Dunn, Robert Andrew / <i>East Tennessee State University, USA</i>	1085
Düzgün, Şebnem / <i>Colorado School of Mines, USA</i>	1623
Ekere, Justina Ngozi / <i>University of Nigeria, Nsukka, Nigeria</i>	1767
Elshater, Abeer / <i>Ain Shams University, Egypt</i>	1196
Erener, Arzu / <i>Kocaeli University, Turkey</i>	1623
Eyinagho, Monday Ofori / <i>Afe Babalola University, Nigeria</i>	967
Fagbohun, Michael O. / <i>Covenant University, Nigeria</i>	1767
Falaki, Samuel Oluwole / <i>Afe Babalola University, Nigeria</i>	967
Fauzi, Syazana / <i>Universiti Brunei Darussalam, Brunei</i>	1860
Favre, Liliana Maria / <i>Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina & Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina</i>	887
Forge, John / <i>University of Sydney, Australia</i>	1424
Fotopoulou, Ourania-Ioanna / <i>Aristotle University of Thessaloniki, Greece</i>	1919
Fridman, Alexander Y. / <i>Institute for Informatics and Mathematical Modeling of RAS, Russia</i>	685

Funes, Ana / <i>Universidad Nacional de San Luis, Argentina</i>	749
Gagliardi, Isabella / <i>IMATI-CNR, Italy</i>	107
Gallerani, Valentina / <i>University of Bologna, Italy</i>	127
Gama, João / <i>INESC TEC, Portugal & FEP, University of Porto, Portugal</i>	675
Gauttier, Stéphanie / <i>Grenoble Ecole de Management, France</i>	1746
Gavurova, Beata / <i>Technical University of Košice, Slovakia</i>	1872
Gegov, Alexander / <i>School of Computing, University of Portsmouth, UK</i>	290, 1271
George, James / <i>Archway Publishing, USA</i>	1550
Gogi, Vyshali J. / <i>Rashtreeya Vidyalaya College of Engineering, India</i>	847
Gómez, Adán A. / <i>Universidad de Córdoba, Colombia</i>	54
Gonçalves, Patrícia C. T. / <i>INESC TEC, University of Porto, Portugal</i>	1827
Goudos, Sotirios K. / <i>Aristotle University of Thessaloniki, Greece</i>	950
Govindarajan, M. / <i>Annamalai University, India</i>	577, 768
Gutiérrez-Artacho, Juncal / <i>Department of Translation and Interpreting, University of Granada, Spain</i>	874
Hable, Asawaree Anand / <i>MAEER's Maharashtra Institute of Pharmacy, Pune, India</i>	1797
Hadad, Graciela Dora Susana / <i>Universidad Nacional del Oeste, Argentina & Universidad de Belgrano, Argentina</i>	1165
Hafi, Ynes / <i>ARSELA, Tunisia</i>	1578
Hasan, Osman / <i>National University of Sciences and Technology, Pakistan</i>	1002
Hight, Craig / <i>University of Newcastle, Australia</i>	1648
Holland, Barbara Jane / <i>Brooklyn Public Library, USA</i>	1952
Hu, Wen-Chen / <i>University of North Dakota, USA</i>	1384
Ibrahim, Adamkolo Mohammed / <i>University of Maiduguri, Nigeria</i>	344
Ilo, Promise Ifeoma / <i>University Library, Pamo University of Medical Sciences, Port Harcourt, Nigeria</i>	1767
Izuagbe, Roland / <i>Covenant University, Nigeria</i>	1767
Jafari, Raheleh / <i>School of Design, University of Leeds, Leeds, UK</i>	290, 1271
Jagdale, Swati Changdeo / <i>School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India</i>	1797
Jamil, Mohammad Izzuddin Mohammed / <i>Universiti Brunei Darussalam, Brunei</i>	1683
Jerónimo, Alba J. / <i>Universidad de Córdoba, Colombia</i>	54
Jiménez-Fanjul, Noelia / <i>Universidad de Córdoba, Spain</i>	624
Joe, Jennifer Ashley Wright / <i>University of Toledo, USA</i>	1757
K., Brindha / <i>Vellore Institute of Technology, India</i>	98
Kapari, Marianna / <i>INFOLYSiS P.C., Greece</i>	924
Kato, Shogo / <i>Tokyo Woman's Christian University, Japan</i>	904
Kato, Yuuki / <i>Sagami Women's University, Japan</i>	904
Khan, Mohammed Naved / <i>Aligarh Muslim University, India</i>	1106
Khoo, Elaine / <i>The University of Waikato, New Zealand</i>	1648
Khosrowshahi, Yasmine / <i>Wake Forest University, USA</i>	449
Korres, George M. / <i>University of Aegean, Greece</i>	1526
Korstanje, Maximiliano Emanuel / <i>University of Palermo, Argentina</i>	1638
Kostic-Ljubisavljevic, Aleksandra / <i>University of Belgrade, Serbia</i>	1051
Koumaras, Harilaos / <i>NCSR "Demokritos", Greece</i>	924
Koumaras, Vaios / <i>INFOLYSiS P.C., Greece</i>	924

Krmela, Jan / <i>Alexander Dubcek University of Trencin, Slovakia</i>	1289
Kubak, Matus / <i>Technical University of Kosice, Slovakia</i>	1872
Kulik, Boris A. / <i>Institute for Problems in Mechanical Engineering of RAS, Russia</i>	685
Kumar, Adesh / <i>University of Petroleum and Energy Studies, India</i>	815
Kumar, Harendra / <i>Gurukula Kangri Vishwavidyalaya, India</i>	233
Leaning, Marcus / <i>University of Winchester, UK</i>	1672
Ledesma, Viviana Alejandra / <i>Universidad Nacional de La Matanza, Argentina</i>	1165
Lee, Wanbil William / <i>The Computer Ethics Society, Hong Kong & Wanbil and Associates, Hong Kong</i>	411
León-Mantero, Carmen / <i>Universidad de Córdoba, Spain</i>	624
Lo Huang, Jose / <i>Simon Bolivar University, Venezuela</i>	207
Lu, Shen / <i>University of South Florida, USA</i>	602
Luk, Sabrina Ching Yuen / <i>Nanyang Technological University, Singapore</i>	398
Lukman, Salihu / <i>University of Hafr Al-Batin, Saudi Arabia</i>	1231, 1362
M. N., Vijayalakshmi / <i>Rashtreeya Vidyalaya College of Engineering, India</i>	847
M., Divyaprabha / <i>Thiagarajar College of Engineering, India</i>	473
M., Thangavel / <i>Thiagarajar College of Engineering, India</i>	473
Maier, Martin / <i>Institut National de la RechercheScientifique (INRS), Canada</i>	27
Majumder, Abhishek / <i>Tripura University, India</i>	1018
Malheiro, Benedita / <i>Polytechnic Institute of Porto, Portugal & INESC TEC, Portugal</i>	675
Maniou, Theodora A. / <i>University of Cyprus, Cyprus</i>	1130
Marshall, Stephen W. / <i>East Tennessee State University, USA</i>	1085
Maz-Machado, Alexander / <i>Universidad de Córdoba, Spain</i>	624
McKelvey, Nigel / <i>Letterkenny Institute of Technology, Ireland</i>	1719
Michailidis, Maria P. / <i>University of Nicosia, Cyprus</i>	1526
Mikavica, Branka / <i>Faculty of Transport and Traffic Engineering, University of Belgrade, Serbia</i> 1051	
Mitra, Ananda / <i>Wake Forest University, USA</i>	449
Mohammed, Idi Dansuleiman / <i>Land, Air, and Water Consulting, Nigeria</i>	1336
Mollah, Md. Awal Hossain / <i>University of Rajshahi, Bangladesh</i>	1481
Monteiro, José Augusto / <i>Polytechnic Institute of Gaya, Portugal</i>	1513
Moura, Ana S. / <i>LAQV-REQUIMTE, University of Porto, Portugal</i>	1827
Mukherjee, Debapriya / <i>University of Calcutta, India</i>	271
Murrar, Abdullah / <i>Indiana University of Pennsylvania, USA</i>	1550
N., Ambika / <i>Department of Computer Applications, SSMRV College, Bangalore, India</i> 500, 514, 528	
Nair, Rajit / <i>Jagran Lakecity University, India</i>	559
Narang, Ritu / <i>University of Lucknow, India</i>	1066
Nayak, Sabyasachi / <i>Confederation of Indian Industry, India</i>	1353
Nechval, Nicholas / <i>University of Latvia, Latvia</i>	701
Nkiko, Christopher / <i>Elizade University, Nigeria</i>	1767
Oke, Isaiah Adesola / <i>Obafemi Awolowo University, Nigeria</i>	1231, 1336, 1362
Oluwaseun, Fehintola Ezekiel / <i>Adeyemi Federal University of Education, Ondo, Nigeria</i>	1362
Olvera-Lobo, María-Dolores / <i>Department of Information and Communication Sciences, University of Granada, Spain</i>	874
Omolade-Lawal, Aanuoluwapo Oluwaseun / <i>University College Hospital, Ibadan, Nigeria</i>	1699

Onuoha Oragwa, Mgbeodichinma Eucharia / <i>Technical University of Mining Bergakademie Freiberg, Germany</i>	1398
Ostroha, Ruslan / <i>Sumy State University, Ukraine</i>	1289
Pandey, Sarvesh / <i>MMM University of Technology, India</i>	800
Papadopoulos, Georgios O. / <i>National and Kapodistrian University of Athens, Greece</i>	155
Papadopoulos, Ioannis / <i>Metropolitan College, Greece</i>	1662
Papadopoulou, Lambrini / <i>Aristotle University of Thessaloniki, Greece</i>	1130
Papaioannou, Angeliki / <i>INFOLYSiS P.C., Greece</i>	924
Paul, Satyam / <i>Department of Engineering Design and Mathematics, University of the West of England, Bristol, UK</i>	290
Pavlov, Yuri P. / <i>Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria</i>	1309
Pires, Carla / <i>CBIOS, Universidade Lusófona, Portugal</i>	1815
Puaschunder, Julia / <i>The New School, USA</i>	428
Rabuzin, Kornelije / <i>Faculty of Organization and Informatics, University of Zagreb, Croatia</i>	778
Radu, Monica Bixby / <i>Southeast Missouri State University, USA</i>	387, 1096
Rao, Bapuji / <i>Indira Gandhi Institute of Technology, Sarang, India</i>	828
Rascão, José Poças / <i>Polytechnic Institute of Setúbal, Portugal</i>	1728
Rashid, Adnan / <i>National University of Sciences and Technology, Pakistan</i>	1002
Rath, Mamata / <i>Birla School of Management, Birla Global University, India</i>	586
Rathore, Maya / <i>Department of Computer Science, Christian Eminent College, Indore, India</i>	190
Razvarz, Sina / <i>Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	290, 1271
Rehioui, Fadoua / <i>University Soultan Moulay Slimane, Morocco</i>	542
Rivera-Trigueros, Irene / <i>Department of Translation and Interpreting, University of Granada, Spain</i>	874
Rodger, James Allen / <i>Indiana University of Pennsylvania, USA</i>	1550
Rook, Alexandria L. / <i>Southeast Missouri State University, USA</i>	387
Routray, Sudhir K. / <i>Bule Hora University, Bule Hora, Ethiopia</i>	913, 939
Rouzbehani, Khadijeh Roya / <i>University of Victoria, Canada</i>	1850
Saito, Nagayuki / <i>International Professional University of Technology in Tokyo, Japan</i>	1537
Sakkas, Christos / <i>INFOLYSiS P.C., Greece</i>	924
Sakpere, Wilson / <i>University of Bologna, Italy</i>	127
Salihu, Lukman / <i>University of Hafr Al-Batin, Saudi Arabia</i>	1336
Sarasa-Cabezuelo, Antonio / <i>Facultad de Informática, Universidad Complutense de Madrid, Spain</i>	791
Saridou, Theodora A. / <i>Aristotle University of Thessaloniki, Greece</i>	1152
Sarmiento, Rui Portocarrero / <i>Statsframe, University of Porto, Portugal</i>	637
Sarp, Gülcan / <i>Süleyman Demirel University, Turkey</i>	1623
Segall, Richard S. / <i>Arkansas State University, USA</i>	602
Sen, Devjani / <i>Algonquin College, Canada</i>	1839
Senthilkumar, Sudha S. / <i>Vellore Institute of Technology, India</i>	98
Sepesy Maučec, Mirjam / <i>University of Maribor, Slovenia</i>	73
Šestak, Martina / <i>Faculty of Organization and Informatics, University of Zagreb, Croatia</i>	778
Shanker, Udai / <i>MMM University of Technology, India</i>	800
Sharma, Paawan / <i>Pandit Deendayal Petroleum University, India</i>	815

Shibuya, Kazuhiko / ROIS, Japan	1469, 1612
Shmelova, Tetiana / National Aviation University, Ukraine	658
Shukla, Vaibhav / Tech Mahindra, Mumbai, India	980
Sikirda, Yuliya / Flight Academy, National Aviation University, Ukraine	658
Singhania, Udit / Vellore Institute of Technology, India	87
Spallone, Roberta / Politecnico di Torino, Italy.....	1181
Sridhar, Gumpeny R. / Endocrine and Diabetes Centre, Visakhapatnam, India.....	1888
Stergiou, Ioannis / INFOLYSiS P.C., Greece.....	924
Sugiyama, Shigeki / Independent Researcher, Japan.....	177, 1944
Suman, Ugrasen / School of Computer Science and Information Technology, Devi Ahilya University, India	190
Syropoulos, Apostolos / Greek Molecular Computing Group, Greece	155, 493, 1118, 1662
Syropoulos-Harissis, Demetrios Georgios / Greek Molecular Computing Group, Greece	493
Tamzini, Khaled / LaREMFQ Laboratory, IHEC of Sousse, University of Sousse, Tunisia.....	1578
Theodoropoulos, George / INFOLYSiS P.C., Greece.....	924
Tiwari, Sonal / University of Lucknow, India	1066
Tripathy, B. K. / Vellore Institute of Technology, India.....	87
Tzanis, George / Aristotle University of Thessaloniki, Greece.....	1919
Ugwu, Cyprian Ifeanyi / University of Nigeria, Nsukka, Nigeria & University of South Africa, South Africa	1767
Umukoro, Immanuel Ovemeso / Lagos Business School, Pan-Atlantic University, Nigeria	1699
Usuki, Kiminori / Tokyo Metropolitan Mizuho Noge High School, Japan	904
Vaidya, Akshat / Vellore Institute of Technology, India	98
Vargas Jarillo, Cristóbal / Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico	1271
Vargas-Bernal, Rafael / Instituto Tecnológico Superior de Irapuato, Mexico	1254
Veglis, Andreas / Aristotle University of Thessaloniki, Greece	1140, 1152
Veloso, Bruno / INESC TEC, Portugal & University Portucalense, Portugal.....	675
Véstias, Mário Pereira / INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal	12, 245, 257
Vlaj, Damjan / University of Maribor, Slovenia.....	138
Voskoglou, Michael / Graduate Technological Educational Institute of Western Greece, Greece..	315
Walczak, Steven / University of South Florida, USA	1901
Weiss, Andrew Philip / California State University, Northridge, USA.....	1782
Wernicke, Immo H. / German Government (retired), Germany.....	1323
Williams, Idongesit / Aalborg University, Denmark	1436
Wiseman, Yair / Bar-Ilan University, Israel.....	1
Wong, Sam / University College Roosevelt, The Netherlands	1408
Wright, Michelle F. / Pennsylvania State University, USA & Masaryk University, USA	356, 374
Yasuda, Gen'ichi / Nagasaki Institute of Applied Science, Japan.....	42
Yu, Szde / Wichita State University, USA.....	333
Zgank, Andrej / University of Maribor, Slovenia.....	138
Zhizhin, Gennadiy Vladimirovich / Skolkovo, Russia.....	730
Zinyama, Tawanda / University of Zimbabwe, Zimbabwe	1497, 1597

Table of Contents by Category

Acknowledgment	liv
-----------------------------	-----

Preface	xlii
----------------------	------

Volume I

Section 1: Artificial Intelligence

Autonomous Vehicles	1
---------------------------	---

Yair Wiseman, Bar-Ilan University, Israel

Convolutional Neural Network	12
------------------------------------	----

Mário Pereira Véstias, INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal

Human-Agent-Robot Teamwork (HART) Over FiWi-Based Tactile Internet Infrastructures.....	27
---	----

Mahfuzulhoq Chowdhury, Institut National de la Recherche Scientifique (INRS), Canada & University of Quebec, Canada

Martin Maier, Institut National de la Recherche Scientifique (INRS), Canada

Embedded Control System Design for Inverted Pendulum Type Mobile Robots Based on High-Level Petri Nets	42
--	----

Gen'ichi Yasuda, Nagasaki Institute of Applied Science, Japan

Semantic and Formal Representation of Cognitive Models for the Metacognitive Architecture CARINA	54
--	----

Alba J. Jerónimo, Universidad de Córdoba, Colombia

María P. Barrera, Universidad de Córdoba, Colombia

Manuel F. Caro, Universidad de Córdoba, Colombia

Adán A. Gómez, Universidad de Córdoba, Colombia

Error Types in Natural Language Processing in Inflectional Languages	73
--	----

Gregor Donaj, University of Maribor, Slovenia

Mirjam Sepesy Maučec, University of Maribor, Slovenia

Text-Based Image Retrieval Using Deep Learning	87
<i>Udit Singhanian, Vellore Institute of Technology, India</i>	
<i>B. K. Tripathy, Vellore Institute of Technology, India</i>	
Dynamic Load Balancing Using Honey Bee Algorithm: Load Balancing.....	98
<i>Sudha S. Senthilkumar, Vellore Institute of Technology, India</i>	
<i>Brindha K., Vellore Institute of Technology, India</i>	
<i>Nitesh Kumar Agrawal, Vellore Institute of Technology, India</i>	
<i>Akshat Vaidya, Vellore Institute of Technology, India</i>	
Unsupervised Automatic Keyphrases Extraction on Italian Datasets.....	107
<i>Isabella Gagliardi, IMATI-CNR, Italy</i>	
<i>Maria Teresa Artese, IMATI-CNR, Italy</i>	
An Intelligent Machine-Driven Perspective to Archaeological Pottery Reassembly	127
<i>Wilson Sakpere, University of Bologna, Italy</i>	
<i>Valentina Gallerani, University of Bologna, Italy</i>	
Acoustic Presence Detection in a Smart Home Environment.....	138
<i>Andrej Zgank, University of Maribor, Slovenia</i>	
<i>Damjan Vlaj, University of Maribor, Slovenia</i>	

Section 2: Computational Science

Sense and Boundaries of Computer Simulations.....	155
<i>Georgios O. Papadopoulos, National and Kapodistrian University of Athens, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
A General-Purpose Uniform Random Number Package	164
<i>E. Jack Chen, BASF Corporation, USA</i>	
Individual Cloud: After Cloud.....	177
<i>Shigeki Sugiyama, Independent Researcher, Japan</i>	
A Collaborative Ranking Approach for Discovery and Selection of Cloud Services	190
<i>Maya Rathore, Department of Computer Science, Christian Eminent College, Indore, India</i>	
<i>Ugrasen Suman, School of Computer Science and Information Technology, Devi Ahilya University, India</i>	
Tuning Drone Data Delivery and Analysis on the Public Cloud.....	207
<i>Jose Lo Huang, Simon Bolivar University, Venezuela</i>	
Improving the Integration of Distributed Applications	217
<i>José Carlos Martins Delgado, Instituto Superior Técnico, Universidade de Lisboa, Portugal</i>	

A Survey of Tasks Scheduling Algorithms in Distributed Computing Systems	233
<i>Nutan Kumari Chauhan, KIET Group of Institutions, India</i>	
<i>Harendra Kumar, Gurukula Kangri Vishwavidyalaya, India</i>	
High-Speed Viterbi Decoder.....	245
<i>Mário Pereira Véstias, INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto</i>	
<i>Politécnico de Lisboa, Portugal</i>	
Field-Programmable Gate Array.....	257
<i>Mário Pereira Véstias, INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto</i>	
<i>Politécnico de Lisboa, Portugal</i>	
Design of a Simple and Low-Cost Calculator in the Laboratory Using FPGA.....	271
<i>Debapriya Mukherjee, University of Calcutta, India</i>	
<i>Kaustav Das, University of Calcutta, India</i>	
<i>Arpita Das, University of Calcutta, India</i>	
Modeling of Uncertain Nonlinear System With Z-Numbers.....	290
<i>Raheleh Jafari, School of Design, University of Leeds, Leeds, UK</i>	
<i>Sina Razvarz, Department of Automatic Control, Centre for Research and Advanced Studies</i>	
<i>of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	
<i>Alexander Gegov, School of Computing, University of Portsmouth, UK</i>	
<i>Satyam Paul, Department of Engineering Design and Mathematics, University of the West of</i>	
<i>England, Bristol, UK</i>	
Use of Grey Numbers for Evaluating a System's Performance Under Fuzzy Conditions.....	315
<i>Michael Voskoglou, Graduate Technological Educational Institute of Western Greece,</i>	
<i>Greece</i>	

Section 3: Cyber Crime, Cyber Bullying, and Digital Terrorism

Cyber Profiling in Criminal Investigation.....	333
<i>Szde Yu, Wichita State University, USA</i>	
Boko Haram's Feminization, Minorization, and Cyberization of Terrorism: Offering the	
Cyberterrorism Diffusion Model as Anti-Bokoharamism Tool.....	344
<i>Adamkolo Mohammed Ibrahim, University of Maiduguri, Nigeria</i>	
Cyberbullying: Definition, Behaviors, Correlates, and Adjustment Problems	356
<i>Michelle F. Wright, Pennsylvania State University, USA & Masaryk University, USA</i>	
What Is Cyberbullying?	374
<i>Michelle F. Wright, Pennsylvania State University, USA & Masaryk University, USA</i>	

How Theoretical Frameworks Inform the Understanding of the Relationship Between Gender and Cyberbullying	387
<i>Monica Bixby Radu, Southeast Missouri State University, USA</i>	
<i>Alexandria L. Rook, Southeast Missouri State University, USA</i>	
The Human Rights-Based Approach to Combat Cyberbullying Against Women and Girls	398
<i>Sabrina Ching Yuen Luk, Nanyang Technological University, Singapore</i>	

Section 4: Cyber Security and Data Protection

Hexa-Dimension Metric, Ethical Matrix, and Cybersecurity	411
<i>Wanbil William Lee, The Computer Ethics Society, Hong Kong & Wanbil and Associates, Hong Kong</i>	
A Utility Theory of Privacy and Information Sharing	428
<i>Julia Puaschunder, The New School, USA</i>	
The 2018 Facebook Data Controversy and Technological Alienation	449
<i>Ananda Mitra, Wake Forest University, USA</i>	
<i>Yasmine Khosrowshahi, Wake Forest University, USA</i>	
Forensic Acquisition Methods for Cloud Computing Environments	462
<i>Diane Barrett, Bloomsburg University of Pennsylvania, USA</i>	
Threats and Vulnerabilities of Mobile Applications	473
<i>Thangavel M., Thiagarajar College of Engineering, India</i>	
<i>Divyaprabha M., Thiagarajar College of Engineering, India</i>	
<i>Abinaya C., Thiagarajar College of Engineering, India</i>	
Web Authorization Protocols	493
<i>Demetrios Georgios Syropoulos-Harissis, Greek Molecular Computing Group, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
Secure and Reliable Knowledge-Based Intrusion Detection Using Mobile Base Stations in Smart Environments	500
<i>Ambika N., Department of Computer Applications, SSMRV College, Bangalore, India</i>	
Improved Cross-Layer Detection and Prevention of Sinkhole Attack in WSN	514
<i>Ambika N., Department of Computer Application, SSMRV College, Bangalore, India</i>	
TDSJ-IoT: Trivial Data Transmission to Sustain Energy From Reactive Jamming Attack in IoT	528
<i>Ambika N., Department of Computer Application, SSMRV College, Bangalore, India</i>	

Section 5: Data Management, Analysis, and Statistics

Software Component Technology: Concepts, Design, and Management Method	542
<i>Fadoua Rehioui, University Soultan Moulay Slimane, Morocco</i>	
An Introduction to Clustering Algorithms in Big Data	559
<i>Rajit Nair, Jagran Lakecity University, India</i>	
<i>Amit Bhagat, Maulana Azad National Institute of India, India</i>	
Challenges in Big Data Analysis	577
<i>M. Govindarajan, Annamalai University, India</i>	
Big Data Analytics and IoT in Smart City Applications.....	586
<i>Mamata Rath, Birla School of Management, Birla Global University, India</i>	
Data Streaming Processing Window Joined With Graphics Processing Units (GPUs).....	602
<i>Shen Lu, University of South Florida, USA</i>	
<i>Richard S. Segall, Arkansas State University, USA</i>	
Statistical Techniques for Research.....	624
<i>Jose Carlos Casas-Rosal, Universidad de Córdoba, Spain</i>	
<i>Carmen León-Mantero, Universidad de Córdoba, Spain</i>	
<i>Noelia Jiménez-Fanjul, Universidad de Córdoba, Spain</i>	
<i>Alexander Maz-Machado, Universidad de Córdoba, Spain</i>	

Volume II

Panel Data: A Case Study Analysis	637
<i>Vera Costa, FEUP, University of Porto, Portugal</i>	
<i>Rui Portocarrero Sarmento, Statsframe, University of Porto, Portugal</i>	
Applications of Decision Support Systems in Aviation.....	658
<i>Tetiana Shmelova, National Aviation University, Ukraine</i>	
<i>Yuliya Sikirda, Flight Academy, National Aviation University, Ukraine</i>	
Classification and Recommendation With Data Streams	675
<i>Bruno Veloso, INESC TEC, Portugal & University Portucalense, Portugal</i>	
<i>João Gama, INESC TEC, Portugal & FEP, University of Porto, Portugal</i>	
<i>Benedita Malheiro, Polytechnic Institute of Porto, Portugal & INESC TEC, Portugal</i>	
N-Tuple Algebra as a Generalized Theory of Relations	685
<i>Boris A. Kulik, Institute for Problems in Mechanical Engineering of RAS, Russia</i>	
<i>Alexander Y. Fridman, Institute for Informatics and Mathematical Modeling of RAS, Russia</i>	
Intelligent Constructing Exact Tolerance Limits for Prediction of Future Outcomes Under Parametric Uncertainty	701
<i>Nicholas Nechval, University of Latvia, Latvia</i>	

The Structure of DNA Taking Into Account the Higher Dimension of Its Components	730
<i>Gennadiy Vladimirovich Zhizhin, Skolkovo, Russia</i>	

Section 6: Data Mining and Databases

Methods and Techniques of Data Mining	749
<i>Ana Funes, Universidad Nacional de San Luis, Argentina</i>	
<i>Aristides Dasso, Universidad Nacional de San Luis, Argentina</i>	
Big Data Mining Algorithms	768
<i>M. Govindarajan, Annamalai University, India</i>	
Graph Database Management Systems: The Past, the Present, and the Future	778
<i>Kornelije Rabuzin, Faculty of Organization and Informatics, University of Zagreb, Croatia</i>	
<i>Martina Šestak, Faculty of Organization and Informatics, University of Zagreb, Croatia</i>	
New Trends in Databases to NonSQL Databases	791
<i>Antonio Sarasa-Cabezuelo, Facultad de Informática, Universidad Complutense de Madrid, Spain</i>	
EDRC: An Early Data Lending-Based Real-Time Commit Protocol	800
<i>Sarvesh Pandey, MMM University of Technology, India</i>	
<i>Udai Shanker, MMM University of Technology, India</i>	
Application of Data Mining in Smart Grid Technology	815
<i>Lipi Chhaya, Researcher, University of Petroleum and Energy Studies, India</i>	
<i>Paawan Sharma, Pandit Deendayal Petroleum University, India</i>	
<i>Adesh Kumar, University of Petroleum and Energy Studies, India</i>	
<i>Govind Bhagwatikar, Sany Wind Energy India Private Limited, India</i>	
N-Clustering of Text Documents Using Graph Mining Techniques	828
<i>Bapuji Rao, Indira Gandhi Institute of Technology, Sarang, India</i>	
Characterization of Elevated Tumor Markers in Diagnosis of HCC Using Data Mining Methods ...	847
<i>Vyshali J. Gogi, Rashtreeya Vidyalaya College of Engineering, India</i>	
<i>Vijayalakshmi M. N., Rashtreeya Vidyalaya College of Engineering, India</i>	
Designing a Concept-Mining Model for the Extraction of Medical Information in Spanish	856
<i>Olga Acosta, Singulartya SpA, Chile</i>	
<i>César Aguilar, Pontificia Universidad Católica de Chile, Chile</i>	

Section 7: Digital and Wireless Communications

Overview of Machine Translation Development	874
<i>Irene Rivera-Trigueros, Department of Translation and Interpreting, University of Granada, Spain</i>	
<i>María-Dolores Olvera-Lobo, Department of Information and Communication Sciences, University of Granada, Spain</i>	
<i>Juncal Gutiérrez-Artacho, Department of Translation and Interpreting, University of Granada, Spain</i>	
Migrating Software Towards Mobile Technologies.....	887
<i>Liliana Maria Favre, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina & Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina</i>	
Interaction Speed as Nonverbal Cues in Text Messaging via Smartphone.....	904
<i>Kiminori Usuki, Tokyo Metropolitan Mizuho Noge High School, Japan</i>	
<i>Shogo Kato, Tokyo Woman's Christian University, Japan</i>	
<i>Yuuki Kato, Sagami Women's University, Japan</i>	
Narrowband Internet of Things.....	913
<i>Sudhir K. Routray, Bule Hora University, Bule Hora, Ethiopia</i>	
Interoperability Provision of IoT Data Protocols on Top of Virtualized Infrastructure.....	924
<i>Vaios Koumaras, INFOLYSiS P.C., Greece</i>	
<i>Marianna Kapari, INFOLYSiS P.C., Greece</i>	
<i>Angeliki Papaioannou, INFOLYSiS P.C., Greece</i>	
<i>George Theodoropoulos, INFOLYSiS P.C., Greece</i>	
<i>Ioannis Stergiou, INFOLYSiS P.C., Greece</i>	
<i>Christos Sakkas, INFOLYSiS P.C., Greece</i>	
<i>Harilaos Koumaras, NCSR "Demokritos", Greece</i>	
Network Science for Communication Engineering	939
<i>Sudhir K. Routray, Bule Hora University, Bule Hora, Ethiopia</i>	
Application of Biogeography-Based Optimization to Antennas and Wireless Communications	950
<i>Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece</i>	
Bits'-Carrying Capacities of Switched Local Area Networks	967
<i>Monday Ofori Eyinagho, Afe Babalola University, Nigeria</i>	
<i>Samuel Oluwale Falaki, Afe Babalola University, Nigeria</i>	
Design and Analysis of AWG and FDL-Based Optical Switch in Data Centre Network.....	980
<i>Vaibhav Shukla, Tech Mahindra, Mumbai, India</i>	
<i>Dilip Kumar Choubey, Vellore Institute of Technology, Vellore, India</i>	

Formal Verification of ZigBee-Based Routing Protocol for Smart Grids	1002
<i>Adnan Rashid, National University of Sciences and Technology, Pakistan</i>	
<i>Osman Hasan, National University of Sciences and Technology, Pakistan</i>	
QoS Provisioning for Multicast Routing and Channel Assignment in Wireless Mesh Network.....	1018
<i>Abira Banik, Tripura University, India</i>	
<i>Abhishek Majumder, Tripura University, India</i>	
Evaluating the Effectiveness of Multi-Web Services in Load Balancing Cluster-Based Web Server	1037
<i>Abhijit Bora, Gauhati University, India</i>	
<i>Tulshi Bezboruah, Gauhati University, India</i>	
Challenges and Opportunities of VLC Application in Intelligent Transportation Systems	1051
<i>Aleksandra Kostic-Ljubisavljevic, University of Belgrade, Serbia</i>	
<i>Branka Mikavica, Faculty of Transport and Traffic Engineering, University of Belgrade, Serbia</i>	

Section 8: Digital Media

Antecedents and Consequences of the Growing Popularity of Digital Consumption	1066
<i>Ritu Narang, University of Lucknow, India</i>	
<i>Sonal Tiwari, University of Lucknow, India</i>	
Television Use and Consumption of Elderly Americans	1085
<i>Robert Andrew Dunn, East Tennessee State University, USA</i>	
<i>Stephen W. Marshall, East Tennessee State University, USA</i>	
Social Media and Social Movements: Strengths, Challenges, and Implications for the Future	1096
<i>Sheldondra J. Brown, Southeast Missouri State University, USA</i>	
<i>Grace M. Babcock, North Carolina State University, USA</i>	
<i>Monica Bixby Radu, Southeast Missouri State University, USA</i>	
Digital Social Networks From a Social Capital Perspective.....	1106
<i>Suparna Dhar, NSHM Knowledge Campus, Kolkata, India</i>	
<i>Indranil Bose, Indian Institute of Management, Kolkata, India</i>	
<i>Mohammed Naved Khan, Aligarh Muslim University, India</i>	
Digital Documents	1118
<i>Anastasios Dimou, Independent Researcher, Athens, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
Digital Media and New Forms of Journalism.....	1130
<i>Lambrini Papadopoulou, Aristotle University of Thessaloniki, Greece</i>	
<i>Theodora A. Maniou, University of Cyprus, Cyprus</i>	

Data Journalism: Definition, Skills, Difficulties, and Perspectives.....	1140
<i>Andreas Veglis, Aristotle University of Thessaloniki, Greece</i>	
<i>Charalampos Bratsas, Aristotle University of Thessaloniki, Greece</i>	

Exploring the Integration of User-Generated Content in Media Organizations Through Participatory Journalism	1152
<i>Theodora A. Saridou, Aristotle University of Thessaloniki, Greece</i>	
<i>Andreas Veglis, Aristotle University of Thessaloniki, Greece</i>	

Section 9: Engineering Science and Technologies

Adapting a Requirements Engineering Process by Key Factors Estimation.....	1165
<i>Graciela Dora Susana Hadad, Universidad Nacional del Oeste, Argentina & Universidad de Belgrano, Argentina</i>	
<i>Jorge Horacio Doorn, Universidad Nacional de La Matanza, Argentina & Universidad Nacional de Tres de Febrero, Argentina</i>	
<i>Viviana Alejandra Ledesma, Universidad Nacional de La Matanza, Argentina</i>	

Digital Tools Aimed to Represent Urban Survey	1181
<i>Cristina Boido, Politecnico di Torino, Italy</i>	
<i>Pia Davico, Politecnico di Torino, Italy</i>	
<i>Roberta Spallone, Politecnico di Torino, Italy</i>	

Addressing the New Pragmatic Methods in Urban Design Discipline	1196
<i>Hisham Abusaada, Housing and Building National Research Center (HBRC), Egypt</i>	
<i>Abeer Elshater, Ain Shams University, Egypt</i>	

Reengineering Legacy Systems Towards New Technologies.....	1214
<i>Djelloul Bouchiha, Center University of Naama, Algeria & EEDIS Lab. UDL-SBA, Algeria</i>	

A Comprehensive Update and Performance Evaluation of Friction Factor Formulae	1231
<i>Salihu Lukman, University of Hafr Al-Batin, Saudi Arabia</i>	
<i>Isaiah Adesola Oke, Obafemi Awolowo University, Nigeria</i>	
<i>Afolabi M. Asani, Ladoke Akintola University, Nigeria</i>	

The Role of Two-Dimensional Materials in Electromagnetic Interference Shielding.....	1254
<i>Rafael Vargas-Bernal, Instituto Tecnológico Superior de Irapuato, Mexico</i>	

Effect of Nanoparticle and Inclination Angle at Thermal Efficiency in Heatpipes	1271
<i>Sina Razvarz, Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	
<i>Raheleh Jafari, School of Design, University of Leeds, Leeds, UK</i>	
<i>Cristóbal Vargas Jarillo, Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	
<i>Alexander Gegov, School of Computing, University of Portsmouth, UK</i>	

Modeling of the Aerodisperse Systems Hydrodynamics in Devices With Directional Motion of the Fluidized Bed	1289
<i>Artem Artyukhov, Sumy State University, Ukraine</i>	
<i>Jan Krmela, Alexander Dubcek University of Trencin, Slovakia</i>	
<i>Nadiia Artyukhova, Sumy State University, Ukraine</i>	
<i>Ruslan Ostroha, Sumy State University, Ukraine</i>	

Volume III

Section 10: Environmental Science and Technologies

Renewable Resources and Value-Based Complex Forest Management	1309
<i>Yuri P. Pavlov, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria</i>	
Agriculture 4.0 and Bioeconomy: Strategies of the European Union and Germany to Promote the Agricultural Sector – Opportunities and Strains of Digitization and the Use of Bio-Based Innovations.....	1323
<i>Immo H. Wernicke, German Government (retired), Germany</i>	
Overview of Solid Waste Management of Healthcare and Related Organisations.....	1336
<i>Isaiah Adesola Oke, Obafemi Awolowo University, Nigeria</i>	
<i>Lukman Salihu, University of Hafr Al-Batin, Saudi Arabia</i>	
<i>Idi Dansuleiman Mohammed, Land, Air, and Water Consulting, Nigeria</i>	
<i>Asani M. Afolabi, Ladoke Akintola University, Nigeria</i>	
Managing Water Resources: Industry Initiative.....	1353
<i>Sabyasachi Nayak, Confederation of Indian Industry, India</i>	
Inhibited Antibiotic-Resistant and Electrochemical Treatment of Pharmaceutical Wastewater	1362
<i>Isaiah Adesola Oke, Obafemi Awolowo University, Nigeria</i>	
<i>Fehintola Ezekiel Oluwaseun, Adeyemi Federal University of Education, Ondo, Nigeria</i>	
<i>Justinah S. Amoko, Adeyemi Federal University of Education, Ondo, Nigeria</i>	
<i>Salihu Lukman, University of Hafr Al-Batin, Saudi Arabia</i>	
<i>Adekunbi Enoch Adedayo, Adeyemi Federal University of Education, Ondo, Nigeria</i>	
Design and Implementation of a Location-Based Service With Emphasis on a Geographical Database.....	1384
<i>Wen-Chen Hu, University of North Dakota, USA</i>	
Climate Change: Inclusion of Gender and Cultural Diversity in Climate Change Actions.....	1398
<i>Mgbeodichinma Eucharia Onuoha Oragwa, Technical University of Mining Bergakadmie Freiberg, Germany</i>	
Gendering Information and Communication Technologies in Climate Change.....	1408
<i>Sam Wong, University College Roosevelt, The Netherlands</i>	

Section 11: Ethics and Responsibility

Science and the Morality of Weapons Research.....	1424
<i>John Forge, University of Sydney, Australia</i>	
Mobilization: Decision Theory	1436
<i>Idongesit Williams, Aalborg University, Denmark</i>	
Moral, Social, and Political Responsibility in the Information Age	1451
<i>Tomas Cahlik, Charles University, Prague, Czech Republic</i>	
The Voice of Social Media, 1997-2018	1461
<i>Bruce Lawrence Cook, Cook Communication, USA</i>	
Breaking Fake News and Verifying Truth	1469
<i>Kazuhiko Shibuya, ROIS, Japan</i>	
Crossfire and Violation of Human Rights in Bangladesh: A Critical Review	1481
<i>Md. Awal Hossain Mollah, University of Rajshahi, Bangladesh</i>	
Politics and Ethics and a National Framework to Combat Corruption in Zimbabwe.....	1497
<i>Tawanda Zinyama, University of Zimbabwe, Zimbabwe</i>	

Section 12: Government, Policy, and Crisis Management

Good Practices in E-Government Accessibility: Lessons From the European Union.....	1513
<i>Fernando Almeida, Polytechnic Institute of Gaya, Portugal</i>	
<i>José Augusto Monteiro, Polytechnic Institute of Gaya, Portugal</i>	
National Systems of Innovation Using an Application on EU Data.....	1526
<i>George M. Korres, University of Aegean, Greece</i>	
<i>Maria P. Michailidis, University of Nicosia, Cyprus</i>	
Comparative Analysis of Applying Behavioral Public Policy to Telecommunication Market by International Organizations.....	1537
<i>Nagayuki Saito, International Professional University of Technology in Tokyo, Japan</i>	
A Good American President	1550
<i>James George, Archway Publishing, USA</i>	
<i>Abdullah Murrar, Indiana University of Pennsylvania, USA</i>	
<i>Pankaj Chaudhary, Indiana University of Pennsylvania, USA</i>	
<i>James Allen Rodger, Indiana University of Pennsylvania, USA</i>	

The Role of DPPs in Promoting Local Government-Citizen Collaboration and Participation: The Case of “Baladiaty”	1578
<i>Khaled Tamzini, LaREMFQ Laboratory, IHEC of Sousse, University of Sousse, Tunisia</i>	
<i>Ynes Hafi, ARSELA, Tunisia</i>	
<i>Achref Ben Ouannes, ARSELA, Tunisia</i>	
<i>Roula Borhani, OMEDRH, Tunis, Tunisia</i>	
Local Government, Decentralisation, Devolution, and Service Delivery in Zimbabwe.....	1597
<i>Tawanda Zinyama, University of Zimbabwe, Zimbabwe</i>	
A Risk Management on Demographic Mobility of Evacuees in Disasters.....	1612
<i>Kazuhiko Shibuya, ROIS, Japan</i>	
GIS-Based Quantitative Landslide Risk Assessment Approach for Property and Life at Bartın Hepler.....	1623
<i>Arzu Erener, Kocaeli University, Turkey</i>	
<i>Gülcan Sarp, Süleyman Demirel University, Turkey</i>	
<i>Şebnem Düzgün, Colorado School of Mines, USA</i>	
Section 13: Human Aspects of Technology	
Technology and the Theory of Apocalypse	1638
<i>Maximiliano Emanuel Korstanje, University of Palermo, Argentina</i>	
Software Literacy as a Vital Digital Literacy in a Software-Saturated World.....	1648
<i>Craig Hight, University of Newcastle, Australia</i>	
<i>Elaine Khoo, The University of Waikato, New Zealand</i>	
ICT Investments on Economic Sectors With International Comparative Advantage and the Diffusion of Prosperity.....	1662
<i>Ioannis Papadopoulos, Metropolitan College, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
Dimensions of the Digital Divide	1672
<i>Marcus Leaning, University of Winchester, UK</i>	
<i>Udo Richard Averweg, eThekweni Municipality, South Africa</i>	
Importance of Digital Literacy and Hindrance Brought About by Digital Divide.....	1683
<i>Mohammad Izzuddin Mohammed Jamil, Universiti Brunei Darussalam, Brunei</i>	
<i>Mohammad Nabil Almunawar, Universiti Brunei Darussalam, Brunei</i>	

Gender Differences in Access to and Use of ICTs in Nigeria.....	1699
<i>Immanuel Ovemeso Umukoro, Lagos Business School, Pan-Atlantic University, Nigeria</i>	
<i>Aanuoluwapo Oluwaseun Omolade-Lawal, University College Hospital, Ibadan, Nigeria</i>	
<i>Samuel Oyelami Babalola, UAC Nigeria Plc, Nigeria</i>	
<i>Kolawole Sunday Akinsumbo, LucrativeTech, Nigeria</i>	
<i>Rashida Mebude Aligwa, Tek Experts Limited, Nigeria</i>	
<i>Balikis Animasaun Abdul-Jeleel, International Institute of Tropical Agriculture (IITA), Nigeria</i>	
How Mobile Technologies Are Leading to Economic Development in Sub-Saharan Africa	1719
<i>Nigel McKelvey, Letterkenny Institute of Technology, Ireland</i>	
<i>Adam Crossan, Letterkenny Institute of Technology, Ireland</i>	
<i>Kevin Curran, Ulster University, UK</i>	

Section 14: Library Science and Research Methods

Informational Competencies	1728
<i>José Poças Rascão, Polytechnic Institute of Setúbal, Portugal</i>	
A Primer on Q-Method and the Study of Technology	1746
<i>Stéphanie Gauttier, Grenoble Ecole de Management, France</i>	
Modernizing the Academic Library.....	1757
<i>Jennifer Ashley Wright Joe, University of Toledo, USA</i>	
Prospects and Challenges of Web 3.0 Technologies Application in the Provision of Library Services.....	1767
<i>Promise Ifeoma Ilo, University Library, Pamo University of Medical Sciences, Port Harcourt, Nigeria</i>	
<i>Christopher Nkiko, Elizade University, Nigeria</i>	
<i>Cyprian Ifeanyi Ugwu, University of Nigeria, Nsukka, Nigeria & University of South Africa, South Africa</i>	
<i>Justina Ngozi Ekere, University of Nigeria, Nsukka, Nigeria</i>	
<i>Roland Izuagbe, Covenant University, Nigeria</i>	
<i>Michael O. Fagbohun, Covenant University, Nigeria</i>	
Massive Digital Libraries (MDLs) and the Impact of Mass-Digitized Book Collections.....	1782
<i>Andrew Philip Weiss, California State University, Northridge, USA</i>	

Section 15: Medical Ethics and Law

Protocol Development in Clinical Trials for Healthcare Management.....	1797
<i>Swati Changdeo Jagdale, School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India</i>	
<i>Asawaree Anand Hable, MAEER's Maharashtra Institute of Pharmacy, Pune, India</i>	
<i>Anuruddha R. Chabukswar, School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India</i>	

A Regulatory and Safety Perspective on Medical Devices	1815
<i>Carla Pires, CBIOS, Universidade Lusófona, Portugal</i>	
<i>Dinah Duarte, CBIOS, Universidade Lusófona, Portugal</i>	
Medical Social Networks, Epidemiology and Health Systems	1827
<i>Patrícia C. T. Gonçalves, INESC TEC, University of Porto, Portugal</i>	
<i>Ana S. Moura, LAQV-REQUIMTE, University of Porto, Portugal</i>	
<i>M. Natália D. S. Cordeiro, LAQV-REQUIMTE, University of Porto, Portugal</i>	
<i>Pedro Campos, INESC TEC, University of Porto, Portugal</i>	
Ambiguities in the Privacy Policies of Common Health and Fitness Apps	1839
<i>Devjani Sen, Algonquin College, Canada</i>	
<i>Rukhsana Ahmed, University at Albany, SUNY, USA</i>	
Generosity in Healthcare Policy Under the Obama Administration: Reflecting Various Dimensions Focused on the ACA	1850
<i>Khadijeh Roya Rouzbehani, University of Victoria, Canada</i>	
Islamic Medical Ethics in Brunei Darussalam Hospitals	1860
<i>Syazana Fauzi, Universiti Brunei Darussalam, Brunei</i>	
Distribution of Selected Health Technology in Regions of Slovakia	1872
<i>Beata Gavurova, Technical University of Košice, Slovakia</i>	
<i>Matus Kubak, Technical University of Kosice, Slovakia</i>	
Section 16: Medical Technologies	
Expanding Scope of Information Technology in Clinical Care	1888
<i>Gumpeny R. Sridhar, Endocrine and Diabetes Centre, Visakhapatnam, India</i>	
Artificial Neural Networks in Medicine: Recent Advances	1901
<i>Steven Walczak, University of South Florida, USA</i>	
Internet of Things and Data Science in Healthcare	1919
<i>George Tzanis, Aristotle University of Thessaloniki, Greece</i>	
<i>Ourania-Ioanna Fotopoulou, Aristotle University of Thessaloniki, Greece</i>	
Internet of Things in Healthcare as an Innovative Form of Personalized Medicine	1933
<i>Ljubica Dikovic, College of Applied Sciences Uzice, Serbia</i>	
Highly Aged and After Cloud	1944
<i>Shigeki Sugiyama, Independent Researcher, Japan</i>	
Gene Editing Technology and Ethical Issues	1952
<i>Barbara Jane Holland, Brooklyn Public Library, USA</i>	
About the Contributors	lvi
Index	xcviii

Table of Contents by Volume

Preface	xlii
----------------------	------

Acknowledgment	liv
-----------------------------	-----

Volume I

Autonomous Vehicles	1
<i>Yair Wiseman, Bar-Ilan University, Israel</i>	

Convolutional Neural Network	12
<i>Mário Pereira Véstias, INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal</i>	

Human-Agent-Robot Teamwork (HART) Over FiWi-Based Tactile Internet Infrastructures.....	27
<i>Mahfuzulhoq Chowdhury, Institut National de la Recherche Scientifique (INRS), Canada & University of Quebec, Canada</i>	
<i>Martin Maier, Institut National de la Recherche Scientifique (INRS), Canada</i>	

Embedded Control System Design for Inverted Pendulum Type Mobile Robots Based on High-Level Petri Nets.....	42
<i>Gen'ichi Yasuda, Nagasaki Institute of Applied Science, Japan</i>	

Semantic and Formal Representation of Cognitive Models for the Metacognitive Architecture CARINA	54
<i>Alba J. Jerónimo, Universidad de Córdoba, Colombia</i>	
<i>María P. Barrera, Universidad de Córdoba, Colombia</i>	
<i>Manuel F. Caro, Universidad de Córdoba, Colombia</i>	
<i>Adán A. Gómez, Universidad de Córdoba, Colombia</i>	

Error Types in Natural Language Processing in Inflectional Languages.....	73
<i>Gregor Donaj, University of Maribor, Slovenia</i>	
<i>Mirjam Sepesy Maučec, University of Maribor, Slovenia</i>	

Text-Based Image Retrieval Using Deep Learning	87
<i>Udit Singhania, Vellore Institute of Technology, India</i>	
<i>B. K. Tripathy, Vellore Institute of Technology, India</i>	

Dynamic Load Balancing Using Honey Bee Algorithm: Load Balancing.....	98
<i>Sudha S. Senthilkumar, Vellore Institute of Technology, India</i>	
<i>Brindha K., Vellore Institute of Technology, India</i>	
<i>Nitesh Kumar Agrawal, Vellore Institute of Technology, India</i>	
<i>Akshat Vaidya, Vellore Institute of Technology, India</i>	
Unsupervised Automatic Keyphrases Extraction on Italian Datasets.....	107
<i>Isabella Gagliardi, IMATI-CNR, Italy</i>	
<i>Maria Teresa Artese, IMATI-CNR, Italy</i>	
An Intelligent Machine-Driven Perspective to Archaeological Pottery Reassembly	127
<i>Wilson Sakpere, University of Bologna, Italy</i>	
<i>Valentina Gallerani, University of Bologna, Italy</i>	
Acoustic Presence Detection in a Smart Home Environment.....	138
<i>Andrej Zgank, University of Maribor, Slovenia</i>	
<i>Damjan Vlaj, University of Maribor, Slovenia</i>	
Sense and Boundaries of Computer Simulations.....	155
<i>Georgios O. Papadopoulos, National and Kapodistrian University of Athens, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
A General-Purpose Uniform Random Number Package	164
<i>E. Jack Chen, BASF Corporation, USA</i>	
Individual Cloud: After Cloud	177
<i>Shigeki Sugiyama, Independent Researcher, Japan</i>	
A Collaborative Ranking Approach for Discovery and Selection of Cloud Services	190
<i>Maya Rathore, Department of Computer Science, Christian Eminent College, Indore, India</i>	
<i>Ugrasen Suman, School of Computer Science and Information Technology, Devi Ahilya University, India</i>	
Tuning Drone Data Delivery and Analysis on the Public Cloud	207
<i>Jose Lo Huang, Simon Bolivar University, Venezuela</i>	
Improving the Integration of Distributed Applications	217
<i>José Carlos Martins Delgado, Instituto Superior Técnico, Universidade de Lisboa, Portugal</i>	
A Survey of Tasks Scheduling Algorithms in Distributed Computing Systems	233
<i>Nutan Kumari Chauhan, KIET Group of Institutions, India</i>	
<i>Harendra Kumar, Gurukula Kangri Vishwavidyalaya, India</i>	
High-Speed Viterbi Decoder.....	245
<i>Mário Pereira Véstias, INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal</i>	

Field-Programmable Gate Array.....	257
<i>Mário Pereira Véstias, INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal</i>	
Design of a Simple and Low-Cost Calculator in the Laboratory Using FPGA.....	271
<i>Debapriya Mukherjee, University of Calcutta, India</i>	
<i>Kaustav Das, University of Calcutta, India</i>	
<i>Arpita Das, University of Calcutta, India</i>	
Modeling of Uncertain Nonlinear System With Z-Numbers.....	290
<i>Raheleh Jafari, School of Design, University of Leeds, Leeds, UK</i>	
<i>Sina Razvarz, Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	
<i>Alexander Gegov, School of Computing, University of Portsmouth, UK</i>	
<i>Satyam Paul, Department of Engineering Design and Mathematics, University of the West of England, Bristol, UK</i>	
Use of Grey Numbers for Evaluating a System's Performance Under Fuzzy Conditions.....	315
<i>Michael Voskoglou, Graduate Technological Educational Institute of Western Greece, Greece</i>	
Cyber Profiling in Criminal Investigation.....	333
<i>Szde Yu, Wichita State University, USA</i>	
Boko Haram's Feminization, Minorization, and Cyberization of Terrorism: Offering the Cyberterrorism Diffusion Model as Anti-Bokoharamism Tool.....	344
<i>Adamkolo Mohammed Ibrahim, University of Maiduguri, Nigeria</i>	
Cyberbullying: Definition, Behaviors, Correlates, and Adjustment Problems	356
<i>Michelle F. Wright, Pennsylvania State University, USA & Masaryk University, USA</i>	
What Is Cyberbullying?	374
<i>Michelle F. Wright, Pennsylvania State University, USA & Masaryk University, USA</i>	
How Theoretical Frameworks Inform the Understanding of the Relationship Between Gender and Cyberbullying	387
<i>Monica Bixby Radu, Southeast Missouri State University, USA</i>	
<i>Alexandria L. Rook, Southeast Missouri State University, USA</i>	
The Human Rights-Based Approach to Combat Cyberbullying Against Women and Girls	398
<i>Sabrina Ching Yuen Luk, Nanyang Technological University, Singapore</i>	
Hexa-Dimension Metric, Ethical Matrix, and Cybersecurity	411
<i>Wanbil William Lee, The Computer Ethics Society, Hong Kong & Wanbil and Associates, Hong Kong</i>	

A Utility Theory of Privacy and Information Sharing.....	428
<i>Julia Puauschunder, The New School, USA</i>	
The 2018 Facebook Data Controversy and Technological Alienation	449
<i>Ananda Mitra, Wake Forest University, USA</i>	
<i>Yasmine Khosrowshahi, Wake Forest University, USA</i>	
Forensic Acquisition Methods for Cloud Computing Environments.....	462
<i>Diane Barrett, Bloomsburg University of Pennsylvania, USA</i>	
Threats and Vulnerabilities of Mobile Applications.....	473
<i>Thangavel M., Thiagarajar College of Engineering, India</i>	
<i>Divyaprabha M., Thiagarajar College of Engineering, India</i>	
<i>Abinaya C., Thiagarajar College of Engineering, India</i>	
Web Authorization Protocols.....	493
<i>Demetrios Georgios Syropoulos-Harissis, Greek Molecular Computing Group, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
Secure and Reliable Knowledge-Based Intrusion Detection Using Mobile Base Stations in Smart Environments	500
<i>Ambika N., Department of Computer Applications, SSMRV College, Bangalore, India</i>	
Improved Cross-Layer Detection and Prevention of Sinkhole Attack in WSN	514
<i>Ambika N., Department of Computer Application, SSMRV College, Bangalore, India</i>	
TDSJ-IoT: Trivial Data Transmission to Sustain Energy From Reactive Jamming Attack in IoT	528
<i>Ambika N., Department of Computer Application, SSMRV College, Bangalore, India</i>	
Software Component Technology: Concepts, Design, and Management Method	542
<i>Fadoua Rehioui, University Souldan Moulay Slimane, Morocco</i>	
An Introduction to Clustering Algorithms in Big Data	559
<i>Rajit Nair, Jagran Lakecity University, India</i>	
<i>Amit Bhagat, Maulana Azad National Institute of India, India</i>	
Challenges in Big Data Analysis	577
<i>M. Govindarajan, Annamalai University, India</i>	
Big Data Analytics and IoT in Smart City Applications.....	586
<i>Mamata Rath, Birla School of Management, Birla Global University, India</i>	
Data Streaming Processing Window Joined With Graphics Processing Units (GPUs).....	602
<i>Shen Lu, University of South Florida, USA</i>	
<i>Richard S. Segall, Arkansas State University, USA</i>	

Statistical Techniques for Research.....	624
<i>Jose Carlos Casas-Rosal, Universidad de Córdoba, Spain</i>	
<i>Carmen León-Mantero, Universidad de Córdoba, Spain</i>	
<i>Noelia Jiménez-Fanjul, Universidad de Córdoba, Spain</i>	
<i>Alexander Maz-Machado, Universidad de Córdoba, Spain</i>	

Volume II

Panel Data: A Case Study Analysis	637
<i>Vera Costa, FEUP, University of Porto, Portugal</i>	
<i>Rui Portocarrero Sarmiento, Statsframe, University of Porto, Portugal</i>	
Applications of Decision Support Systems in Aviation.....	658
<i>Tetiana Shmelova, National Aviation University, Ukraine</i>	
<i>Yuliya Sikirda, Flight Academy, National Aviation University, Ukraine</i>	
Classification and Recommendation With Data Streams	675
<i>Bruno Veloso, INESC TEC, Portugal & University Portucalense, Portugal</i>	
<i>João Gama, INESC TEC, Portugal & FEP, University of Porto, Portugal</i>	
<i>Benedita Malheiro, Polytechnic Institute of Porto, Portugal & INESC TEC, Portugal</i>	
N-Tuple Algebra as a Generalized Theory of Relations	685
<i>Boris A. Kulik, Institute for Problems in Mechanical Engineering of RAS, Russia</i>	
<i>Alexander Y. Fridman, Institute for Informatics and Mathematical Modeling of RAS, Russia</i>	
Intelligent Constructing Exact Tolerance Limits for Prediction of Future Outcomes Under Parametric Uncertainty	701
<i>Nicholas Nechval, University of Latvia, Latvia</i>	
The Structure of DNA Taking Into Account the Higher Dimension of Its Components	730
<i>Gennadiy Vladimirovich Zhizhin, Skolkovo, Russia</i>	
Methods and Techniques of Data Mining.....	749
<i>Ana Funes, Universidad Nacional de San Luis, Argentina</i>	
<i>Aristides Dasso, Universidad Nacional de San Luis, Argentina</i>	
Big Data Mining Algorithms.....	768
<i>M. Govindarajan, Annamalai University, India</i>	
Graph Database Management Systems: The Past, the Present, and the Future	778
<i>Kornelije Rabuzin, Faculty of Organization and Informatics, University of Zagreb, Croatia</i>	
<i>Martina Šestak, Faculty of Organization and Informatics, University of Zagreb, Croatia</i>	
New Trends in Databases to NonSQL Databases	791
<i>Antonio Sarasa-Cabezuelo, Facultad de Informática, Universidad Complutense de Madrid, Spain</i>	

EDRC: An Early Data Lending-Based Real-Time Commit Protocol	800
<i>Sarvesh Pandey, MMM University of Technology, India</i>	
<i>Udai Shanker, MMM University of Technology, India</i>	
Application of Data Mining in Smart Grid Technology	815
<i>Lipi Chhaya, Researcher, University of Petroleum and Energy Studies, India</i>	
<i>Paawan Sharma, Pandit Deendayal Petroleum University, India</i>	
<i>Adesh Kumar, University of Petroleum and Energy Studies, India</i>	
<i>Govind Bhagwatikar, Sany Wind Energy India Private Limited, India</i>	
N-Clustering of Text Documents Using Graph Mining Techniques	828
<i>Bapuji Rao, Indira Gandhi Institute of Technology, Sarang, India</i>	
Characterization of Elevated Tumor Markers in Diagnosis of HCC Using Data Mining Methods ...	847
<i>Vyshali J. Gogi, Rashtreeya Vidyalaya College of Engineering, India</i>	
<i>Vijayalakshmi M. N., Rashtreeya Vidyalaya College of Engineering, India</i>	
Designing a Concept-Mining Model for the Extraction of Medical Information in Spanish	856
<i>Olga Acosta, Singularityta SpA, Chile</i>	
<i>César Aguilar, Pontificia Universidad Católica de Chile, Chile</i>	
Overview of Machine Translation Development	874
<i>Irene Rivera-Trigueros, Department of Translation and Interpreting, University of Granada, Spain</i>	
<i>Maria-Dolores Olvera-Lobo, Department of Information and Communication Sciences, University of Granada, Spain</i>	
<i>Juncal Gutiérrez-Artacho, Department of Translation and Interpreting, University of Granada, Spain</i>	
Migrating Software Towards Mobile Technologies.....	887
<i>Liliana Maria Favre, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina & Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina</i>	
Interaction Speed as Nonverbal Cues in Text Messaging via Smartphone.....	904
<i>Kiminori Usuki, Tokyo Metropolitan Mizuho Noge High School, Japan</i>	
<i>Shogo Kato, Tokyo Woman's Christian University, Japan</i>	
<i>Yuuki Kato, Sagami Women's University, Japan</i>	
Narrowband Internet of Things.....	913
<i>Sudhir K. Routray, Bule Hora University, Bule Hora, Ethiopia</i>	
Interoperability Provision of IoT Data Protocols on Top of Virtualized Infrastructure.....	924
<i>Vaios Koumaras, INFOLYSiS P.C., Greece</i>	
<i>Marianna Kapari, INFOLYSiS P.C., Greece</i>	
<i>Angeliki Papaioannou, INFOLYSiS P.C., Greece</i>	

<i>George Theodoropoulos, INFOLYSiS P.C., Greece</i>	
<i>Ioannis Stergiou, INFOLYSiS P.C., Greece</i>	
<i>Christos Sakkas, INFOLYSiS P.C., Greece</i>	
<i>Harilaos Koumaras, NCSR “Demokritos”, Greece</i>	
Network Science for Communication Engineering	939
<i>Sudhir K. Routray, Bule Hora University, Bule Hora, Ethiopia</i>	
Application of Biogeography-Based Optimization to Antennas and Wireless Communications	950
<i>Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece</i>	
Bits’ -Carrying Capacities of Switched Local Area Networks	967
<i>Monday Ofori Eyinagho, Afe Babalola University, Nigeria</i>	
<i>Samuel Oluwale Falaki, Afe Babalola University, Nigeria</i>	
Design and Analysis of AWG and FDL-Based Optical Switch in Data Centre Network.....	980
<i>Vaibhav Shukla, Tech Mahindra, Mumbai, India</i>	
<i>Dilip Kumar Choubey, Vellore Institute of Technology, Vellore, India</i>	
Formal Verification of ZigBee-Based Routing Protocol for Smart Grids	1002
<i>Adnan Rashid, National University of Sciences and Technology, Pakistan</i>	
<i>Osman Hasan, National University of Sciences and Technology, Pakistan</i>	
QoS Provisioning for Multicast Routing and Channel Assignment in Wireless Mesh Network.....	1018
<i>Abira Banik, Tripura University, India</i>	
<i>Abhishek Majumder, Tripura University, India</i>	
Evaluating the Effectiveness of Multi-Web Services in Load Balancing Cluster-Based Web Server	1037
<i>Abhijit Bora, Gauhati University, India</i>	
<i>Tulshi Bezboruah, Gauhati University, India</i>	
Challenges and Opportunities of VLC Application in Intelligent Transportation Systems	1051
<i>Aleksandra Kostic-Ljubisavljevic, University of Belgrade, Serbia</i>	
<i>Branka Mikavica, Faculty of Transport and Traffic Engineering, University of Belgrade, Serbia</i>	
Antecedents and Consequences of the Growing Popularity of Digital Consumption	1066
<i>Ritu Narang, University of Lucknow, India</i>	
<i>Sonal Tiwari, University of Lucknow, India</i>	
Television Use and Consumption of Elderly Americans	1085
<i>Robert Andrew Dunn, East Tennessee State University, USA</i>	
<i>Stephen W. Marshall, East Tennessee State University, USA</i>	

Social Media and Social Movements: Strengths, Challenges, and Implications for the Future	1096
<i>Sheldondra J. Brown, Southeast Missouri State University, USA</i>	
<i>Grace M. Babcock, North Carolina State University, USA</i>	
<i>Monica Bixby Radu, Southeast Missouri State University, USA</i>	
Digital Social Networks From a Social Capital Perspective.....	1106
<i>Suparna Dhar, NSHM Knowledge Campus, Kolkata, India</i>	
<i>Indranil Bose, Indian Institute of Management, Kolkata, India</i>	
<i>Mohammed Naved Khan, Aligarh Muslim University, India</i>	
Digital Documents	1118
<i>Anastasios Dimou, Independent Researcher, Athens, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
Digital Media and New Forms of Journalism.....	1130
<i>Lambrini Papadopoulou, Aristotle University of Thessaloniki, Greece</i>	
<i>Theodora A. Maniou, University of Cyprus, Cyprus</i>	
Data Journalism: Definition, Skills, Difficulties, and Perspectives.....	1140
<i>Andreas Veglis, Aristotle University of Thessaloniki, Greece</i>	
<i>Charalampos Bratsas, Aristotle University of Thessaloniki, Greece</i>	
Exploring the Integration of User-Generated Content in Media Organizations Through Participatory Journalism	1152
<i>Theodora A. Saridou, Aristotle University of Thessaloniki, Greece</i>	
<i>Andreas Veglis, Aristotle University of Thessaloniki, Greece</i>	
Adapting a Requirements Engineering Process by Key Factors Estimation.....	1165
<i>Graciela Dora Susana Hadad, Universidad Nacional del Oeste, Argentina & Universidad de Belgrano, Argentina</i>	
<i>Jorge Horacio Doorn, Universidad Nacional de La Matanza, Argentina & Universidad Nacional de Tres de Febrero, Argentina</i>	
<i>Viviana Alejandra Ledesma, Universidad Nacional de La Matanza, Argentina</i>	
Digital Tools Aimed to Represent Urban Survey	1181
<i>Cristina Boido, Politecnico di Torino, Italy</i>	
<i>Pia Davico, Politecnico di Torino, Italy</i>	
<i>Roberta Spallone, Politecnico di Torino, Italy</i>	
Addressing the New Pragmatic Methods in Urban Design Discipline	1196
<i>Hisham Abusaada, Housing and Building National Research Center (HBRC), Egypt</i>	
<i>Abeer Elshater, Ain Shams University, Egypt</i>	
Reengineering Legacy Systems Towards New Technologies.....	1214
<i>Djelloul Bouchiha, Center University of Naama, Algeria & EEDIS Lab. UDL-SBA, Algeria</i>	

A Comprehensive Update and Performance Evaluation of Friction Factor Formulae	1231
<i>Salihu Lukman, University of Hafr Al-Batin, Saudi Arabia</i>	
<i>Isaiah Adesola Oke, Obafemi Awolowo University, Nigeria</i>	
<i>Afolabi M. Asani, Ladoke Akintola University, Nigeria</i>	
The Role of Two-Dimensional Materials in Electromagnetic Interference Shielding.....	1254
<i>Rafael Vargas-Bernal, Instituto Tecnológico Superior de Irapuato, Mexico</i>	
Effect of Nanoparticle and Inclination Angle at Thermal Efficiency in Heatpipes	1271
<i>Sina Razvarz, Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	
<i>Raheleh Jafari, School of Design, University of Leeds, Leeds, UK</i>	
<i>Cristóbal Vargas Jarillo, Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico</i>	
<i>Alexander Gegov, School of Computing, University of Portsmouth, UK</i>	
Modeling of the Aerodisperse Systems Hydrodynamics in Devices With Directional Motion of the Fluidized Bed	1289
<i>Artem Artyukhov, Sumy State University, Ukraine</i>	
<i>Jan Krmela, Alexander Dubcek University of Trencin, Slovakia</i>	
<i>Nadiia Artyukhova, Sumy State University, Ukraine</i>	
<i>Ruslan Ostroha, Sumy State University, Ukraine</i>	

Volume III

Renewable Resources and Value-Based Complex Forest Management	1309
<i>Yuri P. Pavlov, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria</i>	
Agriculture 4.0 and Bioeconomy: Strategies of the European Union and Germany to Promote the Agricultural Sector – Opportunities and Strains of Digitization and the Use of Bio-Based Innovations.....	1323
<i>Immo H. Wernicke, German Government (retired), Germany</i>	
Overview of Solid Waste Management of Healthcare and Related Organisations.....	1336
<i>Isaiah Adesola Oke, Obafemi Awolowo University, Nigeria</i>	
<i>Lukman Salihu, University of Hafr Al-Batin, Saudi Arabia</i>	
<i>Idi Dansuleiman Mohammed, Land, Air, and Water Consulting, Nigeria</i>	
<i>Asani M. Afolabi, Ladoke Akintola University, Nigeria</i>	
Managing Water Resources: Industry Initiative.....	1353
<i>Sabyasachi Nayak, Confederation of Indian Industry, India</i>	
Inhibited Antibiotic-Resistant and Electrochemical Treatment of Pharmaceutical Wastewater	1362
<i>Isaiah Adesola Oke, Obafemi Awolowo University, Nigeria</i>	

Fehintola Ezekiel Oluwaseun, Adeyemi Federal University of Education, Ondo, Nigeria
Justinah S. Amoko, Adeyemi Federal University of Education, Ondo, Nigeria
Salihu Lukman, University of Hafr Al-Batin, Saudi Arabia
Adekunbi Enoch Adedayo, Adeyemi Federal University of Education, Ondo, Nigeria

Design and Implementation of a Location-Based Service With Emphasis on a Geographical Database	1384
<i>Wen-Chen Hu, University of North Dakota, USA</i>	
Climate Change: Inclusion of Gender and Cultural Diversity in Climate Change Actions.....	1398
<i>Mgbeodichinma Eucharia Onuoha Oragwa, Technical University of Mining Bergakademie Freiberg, Germany</i>	
Gendering Information and Communication Technologies in Climate Change.....	1408
<i>Sam Wong, University College Roosevelt, The Netherlands</i>	
Science and the Morality of Weapons Research.....	1424
<i>John Forge, University of Sydney, Australia</i>	
Mobilization: Decision Theory	1436
<i>Idongesit Williams, Aalborg University, Denmark</i>	
Moral, Social, and Political Responsibility in the Information Age	1451
<i>Tomas Cahlik, Charles University, Prague, Czech Republic</i>	
The Voice of Social Media, 1997-2018	1461
<i>Bruce Lawrence Cook, Cook Communication, USA</i>	
Breaking Fake News and Verifying Truth	1469
<i>Kazuhiko Shibuya, ROIS, Japan</i>	
Crossfire and Violation of Human Rights in Bangladesh: A Critical Review	1481
<i>Md. Awal Hossain Mollah, University of Rajshahi, Bangladesh</i>	
Politics and Ethics and a National Framework to Combat Corruption in Zimbabwe.....	1497
<i>Tawanda Zinyama, University of Zimbabwe, Zimbabwe</i>	
Good Practices in E-Government Accessibility: Lessons From the European Union.....	1513
<i>Fernando Almeida, Polytechnic Institute of Gaya, Portugal</i> <i>José Augusto Monteiro, Polytechnic Institute of Gaya, Portugal</i>	
National Systems of Innovation Using an Application on EU Data.....	1526
<i>George M. Korres, University of Aegean, Greece</i> <i>Maria P. Michailidis, University of Nicosia, Cyprus</i>	

Comparative Analysis of Applying Behavioral Public Policy to Telecommunication Market by International Organizations.....	1537
<i>Nagayuki Saito, International Professional University of Technology in Tokyo, Japan</i>	
A Good American President	1550
<i>James George, Archway Publishing, USA</i>	
<i>Abdullah Murrar, Indiana University of Pennsylvania, USA</i>	
<i>Pankaj Chaudhary, Indiana University of Pennsylvania, USA</i>	
<i>James Allen Rodger, Indiana University of Pennsylvania, USA</i>	
The Role of DPPs in Promoting Local Government-Citizen Collaboration and Participation: The Case of “Baladiaty”.....	1578
<i>Khaled Tamzini, LaREMFQ Laboratory, IHEC of Sousse, University of Sousse, Tunisia</i>	
<i>Ynes Hafi, ARSELA, Tunisia</i>	
<i>Achref Ben Ouannes, ARSELA, Tunisia</i>	
<i>Roula Borhani, OMEDRH, Tunis, Tunisia</i>	
Local Government, Decentralisation, Devolution, and Service Delivery in Zimbabwe.....	1597
<i>Tawanda Zinyama, University of Zimbabwe, Zimbabwe</i>	
A Risk Management on Demographic Mobility of Evacuees in Disasters.....	1612
<i>Kazuhiko Shibuya, ROIS, Japan</i>	
GIS-Based Quantitative Landslide Risk Assessment Approach for Property and Life at Bartın Hepler.....	1623
<i>Arzu Erener, Kocaeli University, Turkey</i>	
<i>Gülcan Sarp, Süleyman Demirel University, Turkey</i>	
<i>Şebnem Düzgün, Colorado School of Mines, USA</i>	
Technology and the Theory of Apocalypse	1638
<i>Maximiliano Emanuel Korstanje, University of Palermo, Argentina</i>	
Software Literacy as a Vital Digital Literacy in a Software-Saturated World.....	1648
<i>Craig Hight, University of Newcastle, Australia</i>	
<i>Elaine Khoo, The University of Waikato, New Zealand</i>	
ICT Investments on Economic Sectors With International Comparative Advantage and the Diffusion of Prosperity.....	1662
<i>Ioannis Papadopoulos, Metropolitan College, Greece</i>	
<i>Apostolos Syropoulos, Greek Molecular Computing Group, Greece</i>	
Dimensions of the Digital Divide	1672
<i>Marcus Leaning, University of Winchester, UK</i>	
<i>Udo Richard Averweg, eThekweni Municipality, South Africa</i>	

Importance of Digital Literacy and Hindrance Brought About by Digital Divide.....	1683
<i>Mohammad Izzuddin Mohammed Jamil, Universiti Brunei Darussalam, Brunei</i>	
<i>Mohammad Nabil Almunawar, Universiti Brunei Darussalam, Brunei</i>	
Gender Differences in Access to and Use of ICTs in Nigeria.....	1699
<i>Immanuel Ovemeso Umukoro, Lagos Business School, Pan-Atlantic University, Nigeria</i>	
<i>Aanuoluwapo Oluwaseun Omolade-Lawal, University College Hospital, Ibadan, Nigeria</i>	
<i>Samuel Oyelami Babalola, UAC Nigeria Plc, Nigeria</i>	
<i>Kolawole Sunday Akinsumbo, LucrativeTech, Nigeria</i>	
<i>Rashida Mebude Aligwa, Tek Experts Limited, Nigeria</i>	
<i>Balikis Animasaun Abdul-Jeleel, International Institute of Tropical Agriculture (IITA), Nigeria</i>	
How Mobile Technologies Are Leading to Economic Development in Sub-Saharan Africa	1719
<i>Nigel McKelvey, Letterkenny Institute of Technology, Ireland</i>	
<i>Adam Crossan, Letterkenny Institute of Technology, Ireland</i>	
<i>Kevin Curran, Ulster University, UK</i>	
Informational Competencies	1728
<i>José Poças Rascão, Polytechnic Institute of Setúbal, Portugal</i>	
A Primer on Q-Method and the Study of Technology	1746
<i>Stéphanie Gauttier, Grenoble Ecole de Management, France</i>	
Modernizing the Academic Library.....	1757
<i>Jennifer Ashley Wright Joe, University of Toledo, USA</i>	
Prospects and Challenges of Web 3.0 Technologies Application in the Provision of Library Services.....	1767
<i>Promise Ifeoma Ilo, University Library, Pamo University of Medical Sciences, Port Harcourt, Nigeria</i>	
<i>Christopher Nkiko, Elizade University, Nigeria</i>	
<i>Cyprian Ifeanyi Ugwu, University of Nigeria, Nsukka, Nigeria & University of South Africa, South Africa</i>	
<i>Justina Ngozi Ekere, University of Nigeria, Nsukka, Nigeria</i>	
<i>Roland Izuagbe, Covenant University, Nigeria</i>	
<i>Michael O. Fagbohun, Covenant University, Nigeria</i>	
Massive Digital Libraries (MDLs) and the Impact of Mass-Digitized Book Collections.....	1782
<i>Andrew Philip Weiss, California State University, Northridge, USA</i>	

Protocol Development in Clinical Trials for Healthcare Management.....	1797
<i>Swati Changdeo Jagdale, School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India</i>	
<i>Asawaree Anand Hable, MAEER's Maharashtra Institute of Pharmacy, Pune, India</i>	
<i>Anuruddha R. Chabukswar, School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India</i>	
A Regulatory and Safety Perspective on Medical Devices.....	1815
<i>Carla Pires, CBIOS, Universidade Lusófona, Portugal</i>	
<i>Dinah Duarte, CBIOS, Universidade Lusófona, Portugal</i>	
Medical Social Networks, Epidemiology and Health Systems.....	1827
<i>Patrícia C. T. Gonçalves, INESC TEC, University of Porto, Portugal</i>	
<i>Ana S. Moura, LAQV-REQUIMTE, University of Porto, Portugal</i>	
<i>M. Natália D. S. Cordeiro, LAQV-REQUIMTE, University of Porto, Portugal</i>	
<i>Pedro Campos, INESC TEC, University of Porto, Portugal</i>	
Ambiguities in the Privacy Policies of Common Health and Fitness Apps	1839
<i>Devjani Sen, Algonquin College, Canada</i>	
<i>Rukhsana Ahmed, University at Albany, SUNY, USA</i>	
Generosity in Healthcare Policy Under the Obama Administration: Reflecting Various Dimensions Focused on the ACA.....	1850
<i>Khadijeh Roya Rouzbehani, University of Victoria, Canada</i>	
Islamic Medical Ethics in Brunei Darussalam Hospitals	1860
<i>Syazana Fauzi, Universiti Brunei Darussalam, Brunei</i>	
Distribution of Selected Health Technology in Regions of Slovakia	1872
<i>Beata Gavurova, Technical University of Košice, Slovakia</i>	
<i>Matus Kubak, Technical University of Kosice, Slovakia</i>	
Expanding Scope of Information Technology in Clinical Care.....	1888
<i>Gumpeny R. Sridhar, Endocrine and Diabetes Centre, Visakhapatnam, India</i>	
Artificial Neural Networks in Medicine: Recent Advances.....	1901
<i>Steven Walczak, University of South Florida, USA</i>	
Internet of Things and Data Science in Healthcare	1919
<i>George Tzanis, Aristotle University of Thessaloniki, Greece</i>	
<i>Ourania-Ioanna Fotopoulou, Aristotle University of Thessaloniki, Greece</i>	
Internet of Things in Healthcare as an Innovative Form of Personalized Medicine.....	1933
<i>Ljubica Dikovic, College of Applied Sciences Uzice, Serbia</i>	

Highly Aged and After Cloud	1944
<i>Shigeki Sugiyama, Independent Researcher, Japan</i>	
Gene Editing Technology and Ethical Issues.....	1952
<i>Barbara Jane Holland, Brooklyn Public Library, USA</i>	
About the Contributors	lvi
Index	xcviii

Preface

As our world continues its inexorable progression, advancements in all areas of science particularly in the field of information science and technology are rapidly revolutionizing the way the human population interacts, conducts business, governs, and maintains health and well-being. A driving force across all fields and industries continues to be technological innovations which constantly disrupt contemporary theories, methods, and practices, and persistently replaces them with more effective and efficient applications. This has been true more than ever during the recent global crises impacting all facets of life and business.

Understanding information science and technology (IS&T) and its most recent advancements is paramount to improving knowledge and expanding the way human minds think, subsist, and evolve, providing mankind with so many new methods of information and knowledge management. Obviously, the rapid rate of innovations and expansion in the utilization and management of information science and technologies requires a profound understanding of every aspect of knowledge related to the latest research and discoveries within this very important field of study.

To examine the evolution of this publication from its first inaugural release to the present can provide a good understanding of research progressions and the overall value of this latest edition. The first edition of the *Encyclopedia of Information Science and Technology (First Edition)* was released in early 2005 and was comprised of five volumes of invaluable knowledge on breakthroughs, discoveries, and authoritative research results in technological advancements. It positioned itself as the first of its kind in reference publications and supplied researchers with a definitive one-stop reference source. Following the success of the first edition, the *Encyclopedia of Information Science and Technology (Second Edition)* was released in late 2008. It included eight volumes of updated research on crucial topics previously covered in the first edition, as well as an examination of new developments, technologies, and areas of research, allowing the second edition to surpass the first in terms of timeliness, comprehensiveness, and critical acclaim.

Seven years later, the field of information science and technology had progressed so rapidly that ten volumes of completely new content uncovering the most current research findings related to technological, organizational, and managerial issues, challenges, trends, and applications was needed. All-encompassing and lauded as an essential reference for any academic library, the *Encyclopedia of Information Science and Technology (Third Edition)* bridged existing gaps in available references by covering paradigms that were shaping the ever-changing research, theory, and discovery of IS&T.

In 2018, technology had become so ubiquitous and relied upon so heavily by society, that a new edition of the *Encyclopedia of Information Science and Technology (Fourth Edition)* was not only warranted, but demanded by researchers looking to remain up-to-date in a field that had grown by leaps

and bounds in only three short years. Comprised once again of ten volumes and sharing expert research from contributions obtained worldwide, this well-established reference source investigated the most forward-thinking and diverse findings on the impact of information science and new technologies in modern settings that cross all sectors of society, including education, healthcare, government, engineering, business, and physical sciences.

As technology continued to advance and disrupt society at a rapid pace, it was observed that once again, an updated edition in this expansive field was needed. However, this time, the advancements made to businesses and organizations were so extensive, and the content that was acquired from prominent researchers across various areas was so great, it was decided that the encyclopedia would need to be split into two different encyclopedias. A smaller and more precise three-volume *Encyclopedia of Information Science and Technology (Fifth Edition)* that centers on the newest technological innovations and their applications in fields such as medicine, library science, security, and government and another larger five-volume encyclopedia, the *Encyclopedia of Organizational Knowledge, Administration, and Technologies*, that is focused primarily on business and organizational aspects, with both of these encyclopedias offering a greater overall level of accessibility and focus to the readership.

The three-volume *Encyclopedia of Information Science and Technology (Fifth Edition)* is composed of 136 expert peer-reviewed chapters that provide a balanced representation of concepts and issues from researchers around the world. For this edition and all that preceded it, these specially invited researchers were asked to submit proposals describing the topic and scope of their chapters. Each submitted chapter was forwarded to at least three expert external reviewers on a double-blind peer review basis. Only submissions with strong and favorable reviews were chosen as chapters for this edition. In some cases, submissions were sent back for several revisions prior to final acceptance. The goal was to assemble the best minds in the information science and technology field from all over the world to contribute chapters to this encyclopedia and to apply the highest level of quality in selecting chapters for inclusion.

This edition of the *Encyclopedia of Information Science and Technology* only contains previously unpublished research on the most emerging insights in the area of information science and technology, exploring in depth the new methods and processes for understanding modern technological tools and their applications as well as emerging theories and ethical controversies. Academicians, professionals, researchers, librarians, policy makers, and students are all targeted beneficiaries of this extensive encyclopedia, especially those working in the fields of computer science, engineering, library science, medicine, medical technologies, environmental science, ethics, data science, security and forensics, media and communications, government, crisis management, and the social sciences, to name a few. With its vast coverage of conceptual, methodological, and technical aspects of IS&T, including but not limited to such cutting-edge topics as gene editing, the Internet of Things, social media, fake news, natural language processing, digital consumption, Big Data, smart cities, cyberbullying, autonomous vehicles, deep learning, algorithms, cyberterrorism, machine learning, and information privacy, the chapters within this publication are sure to provide readers the tools necessary for further research and discovery in their respective industries and/or fields.

The *Encyclopedia of Information Science and Technology, Fifth Edition* is organized into sixteen sections that provide comprehensive coverage of important topics. These sections are: 1) Artificial Intelligence; 2) Computational Science; 3) Cyber Crime, Cyber Bullying, and Digital Terrorism; 4) Cyber Security and Data Protection; 5) Data Management, Analysis, and Statistics; 6) Data Mining and Databases; 7) Digital and Wireless Communications; 8) Digital Media; 9) Engineering Science and Technologies; 10) Environmental Science and Technologies; 11) Ethics and Social Responsibility; 12)

Government, Policy, and Crisis Management; 13) Human Aspects of Technology; 14) Library Science and Research Methods; 15) Medical Ethics and Law; and 16) Medical Technologies.

Each of the focused sections include diverse chapters centered around a common theme, with each of the chapters then arranged by relevance to the theme and to each other. Two different tables of contents support the reader with navigating the contents by volume and alphabetically. Also, technical, and managerial terms and their definitions are offered in addition to thorough reference sections.

Below are just a few highlighted chapter contributions within each of the sections:

SECTION 1: ARTIFICIAL INTELLIGENCE

This section opens this extensive reference source by highlighting some of the latest advancements taking place in artificial intelligence and its applications across a variety of industries. The opening chapter of this section is titled, “**Autonomous Vehicles,**” authored by Prof. Yair Wiseman of Bar-Ilan University, Israel, which discusses improvements made to modern vehicles that allows them to be converted to entirely automatic systems. Another chapter in this section, “**Human-Agent-Robot Teamwork (HART) over FiWi Based Tactile Internet Infrastructures,**” authored by Mr. Chowdhury Mahfuzulhoq and Dr. Martin Maier, of the Institut National de la Recherche Scientifique, Canada, discusses related work and open challenges for latency-sensitive HART task. The next chapter, “**Embedded Control System Design for Inverted Pendulum Type Mobile Robots Based on High-Level Petri Nets,**” authored by Prof. Gen’ichi Yasuda of Nagasaki Institute of Applied Science, Japan, examines a unified method for designing nonlinear motion control software of traveling control by non-holonomic two-wheeled inverted pendulum mobile robots, where self-balancing control is designed with a backstepping approach is presented. Another chapter within this section, “**Text Based Image Retrieval Using Deep Learning,**” authored by Mr. Udit Singhania and Prof. B Tripathy from Vellore Institute of Technology, India, discusses recurrent neural nets, deep restricted Boltzmann machines, and general adversarial nets and how they revolutionize the field of information retrieval. Concluding this section is the chapter, “**Acoustic Presence Detection in a Smart Home Environment,**” authored by Profs. Andrej Zgank and Damjan Vlaj of the University of Maribor, Slovenia. This chapter presents information on acoustic presence detection that can be used to optimize the smart home human-computer interface with personalization and customization functionalities.

SECTION 2: COMPUTATIONAL SCIENCE

This section includes chapters that investigate recent developments within mathematical models and their ability to solve scientific problems. This section begins with the chapter, “**Sense and Boundaries of Computer Simulations,**” authored by Dr. Apostolos Syropoulos from the Greek Molecular Computing Group, Greece, which examines why computer simulation boundaries happen, how they can be solved, and whether the “ultimate” goal of simulating the whole universe is feasible. In a following chapter, “**A General-Purpose Uniform Random Number Package,**” authored by Dr. E. Jack Chen of the BASF Corporation, United States, examines generating sequences of pseudorandom numbers as a fundamental part of computer simulation systems. In another chapter, “**Tuning Drone Data Delivery and Analysis on the Public Cloud,**” authored by Prof. Jose Lo Huang from Simon Bolivar University,

Venezuela, a set of self-developed software written in Python, C, and Bash is used to improve the speed of data transmission and analysis of drone generated videos, taken in eight different cities in America, to the public cloud of Amazon Web Services. A following chapter, **“Modeling of Uncertain Nonlinear System with Z-numbers,”** authored by Prof. Raheleh Jafari, from the University of Leeds, United Kingdom, Dr. Sina Razvarz of CINVESTAV-IPN (National Polytechnic Institute), Mexico, Prof. Alexander Gegov from University of Portsmouth, United Kingdom, and Prof. Satyam Paul of Orebro University, Sweden, examines how fuzzy equations with Z-number coefficients are used in order to model the fuzzy nonlinear systems. In the next and final chapter in this section, **“Use of Grey Numbers for Evaluating a System’s Performance under Fuzzy Conditions,”** authored by Prof. Michael Voskoglou of the Graduate Technological Educational Institute (T.E.I.), Greece, a method using Grey Numbers as tools is developed for assessing a system’s mean performance, which is useful when utilizing qualitative grades and not numerical scores for this purpose.

SECTION 3: CYBER CRIME, CYBER BULLYING, AND DIGITAL TERRORISM

This section presents a collection of chapters that define cyberbullying and cyber terrorism, exploring the consequences of these online attacks and crimes, and suggests potential remedies in combating this phenomenon. The first chapter opening this section, **“Cyber Profiling in Criminal Investigation,”** authored by Prof. Szde Yu of Wichita State University, United States, introduces what cyber profiling is and how it can help with criminal investigation. A following chapter, **“Cyberbullying,”** authored by Prof. Michelle Wright of Pennsylvania State University, United States, examines and defines cyberbullying among children and adolescents by utilizing multidisciplinary qualitative, quantitative, and mixed-design research methodologies. Another chapter within this section, **“How Theoretical Frameworks Inform the Understanding of the Relationship Between Gender and Cyberbullying,”** authored by Profs. Monica Radu and Alexandria Rook of Southeast Missouri State University, United States, uses an interdisciplinary approach to examine gender differences in cyberbullying and the implications this has for future research and policy development. Concluding this section is the final chapter, **“The Human Rights Based Approach to Combat Cyberbullying Against Women and Girls,”** authored by Prof. Yuen Luk of Nanyang Technological University, Singapore, which explores the adoption of a human rights based approach to fighting against cyberbullying and encouraging the rights to privacy and freedom of expression online for women.

SECTION 4: CYBER SECURITY AND DATA PROTECTION

This section discusses the dangers of cyber-attacks by offering methods and frameworks that can be used to improve cyber security and ensure the protection of data and information online. Opening this section is the chapter, **“Hexa-Dimension Metric, Ethical Matrix, and Cybersecurity,”** authored by Prof. Wanbil Lee of The Computer Ethics Society and Wanbil & Associates, Hong Kong, which develops an ethics-based framework as a remedy to mitigate hacking in cyberspace and illustrates the tools of the Ethical Matrix and Hexa-dimension Metric using the Octopus Saga. The next chapter, **“A Utility Theory of Privacy and Information Sharing,”** authored by Prof. Julia Puaschunder of The New School, United States, covers the problem of individuals’ decisions to share information about themselves on

social media and how big data administrators reap the benefits from putting data together over time. A following chapter, **“The 2018 Facebook Data Controversy and Technological Alienation”**, authored by Profs. Ananda Mitra and Yasmine Khosrowshahi of Wake Forest University, United States, uses the issues surrounding the controversy as a point of departure to explore two related issues – first the way in which the “technologically alienated” individual makes technological decisions without being mindful of the consequences of the technological choices and secondly the way in which the alienation can be utilized to obtain detailed information from the structured and unstructured data that makes up the gigantic corpus of Big Data. Another chapter is, **“Forensic Acquisition Methods for Cloud Computing Environments,”** authored by Prof. Diane Barrett of Bloomsburg University of Pennsylvania, United States. This chapter explores current forensic acquisition processes, how current processes need to be modified for cloud investigations, and what new acquisition methods can help garner evidence from a cloud computing-based environment. The final chapter in this section, **“TDSJ-IoT - Trivial Data Transmission to Sustain Energy from Reactive Jamming Attack in IoT,”** authored by Prof. Ambika Nagaraj of the SSMRV Degree College, India, proposes a method to minimize jamming attacks in the Internet of Things by increasing the reliability of communication, increasing overall effectiveness, and minimizing the energy needed to combat attacks.

SECTION 5: DATA MANAGEMENT, ANALYSIS, AND STATISTICS

This section provides many chapters on the modern utilization of statistical models in assessing the utility and usefulness of data and identifying new approaches and software that can best be used for data management. The first chapter in this section, **“Software Component Technology,”** authored by Prof. Fadoua Rehioui of the University Soultan Moulay Slimane, Morocco, discusses the development of information systems, by their data, and the resultant knowledge by proposing a component identification approach and the management component for data management. In a following chapter, **“Data Streaming Processing Window Join with Graphics Processing Units (GPUs),”** authored by Prof. Shen Lu of the University of South Florida, United States and Prof. Richard Segall from Arkansas State University, United States, covers research that discusses the continuous case of big data often called “Data Streaming” and utilizes the algorithmic side of data stream processing. The next chapter, **“Statistical Techniques for Research,”** authored by Profs. Jose Casas-Rosal, Carmen León-Mantero, Noelia Jiménez-Fanjul, and Alexander Maz-Machado of the University of Córdoba, Spain, creates a classification of the main statistical techniques used in different research fields, adds a brief definition of them and specifies their utility, the data hypothesis needed for their use, and software required. Another chapter within this section, **“Classification and Recommendation with Data Streams,”** authored by Prof. Bruno Veloso of University Portucalense, Portugal, Prof. João Gama of INESC TEC - FEP, Portugal, and Prof. Benedita Malheiro from the Polytechnic Institute of Porto, Portugal, addresses data stream knowledge processing, covers classification, recommendation and evaluation, describes existing algorithms/techniques, and identifies open challenges. The concluding chapter in this section, **“The Structure of DNA Taking into Account the Higher Dimension of Its Components,”** authored by Prof. Gennadiy Zhizhin of Skolkovo, Russian Federation, shows representations and images of the higher dimensionality of the constituent DNA molecules (D - ribose and phosphoric acid ion) described mathematically.

SECTION 6: DATA MINING AND DATABASES

This section features chapters that investigate data mining, including the tasks and techniques associated, as well as exploring more advanced methodologies in managing database systems more effectively and efficiently. Opening this section is the chapter, **“Methods and Techniques of Data Mining,”** authored by Profs. Ana Funes and Aristides Dasso of the Universidad Nacional de San Luis, Argentina. This chapter compares data mining with other disciplines while also providing an overview and classification of data mining tasks and methods complete with issues associated to the discipline and future research directions. The next chapter in this section, **“Graph Database Management Systems – the Past, the Present and the Future,”** authored by Profs. Kornelije Rabuzin and Martina Šestak of the University of Zagreb, Faculty of Organization and Informatics, Croatia, gives a timeline overview of developed graph data storage solutions in order to gain insight into past, present, and future trends of GDBMSs. The next chapter, **“New Trends in Databases to NonSQL Databases,”** authored by Prof. Antonio Sarasa-Cabezuelo of Universidad Complutense de Madrid, Spain, introduces the NonSQL databases currently being used as an alternative to solving the needs raised in the context of big data without the limitations of relational databases. Another chapter in this section, **“Application of Data Mining in Smart Grid Technology,”** authored by Profs. Lipi Chhaya, and Adesh Kumar of UPES, India, Prof. Paawan Sharma of PDP, India, and Prof. Govind Bhagwatikar from Sany Wind Energy India Private Limited, India, explains how data mining techniques can be an effective solution for operation and management in smart grids. Finishing out the section is the chapter, **“Designing a Concept Mining Model for the Extraction of Medical Information in Spanish,”** authored by Prof. Olga Acosta, Singularityta SpA, Chile and Prof. César Aguilar, Pontificia Universidad Católica de Chile, Chile, which sketches the development of a method for mining concepts applied on medical corpora in Spanish, particularly the extraction of specialized terms and the identification of lexical-semantic relationships, in concrete hyponymy/hypernymy and meronymy.

SECTION 7: DIGITAL AND WIRELESS COMMUNICATIONS

This section features chapters that define the many modes of digital and wireless communications and presents a variety of methods, approaches, and solutions for optimizing its usage. Starting this section is the chapter, **“Overview of Machine Translation Development,”** authored by Profs. Irene Rivera-Trigueros, María-Dolores Olvera-Lobo, and Gutiérrez-Artacho from the University of Granada, Spain. This chapter provides an overview of Machine Translation (MT) development since its early beginnings in the late 1950s until present day and reviews the evolution of MT techniques starting from traditional approaches. A following chapter is, **“Interaction Speed as Nonverbal Cues in Text Messaging Via Smartphone,”** authored by Mr. Kiminori Usuki of Tokyo Metropolitan Mizuho Noge High School, Japan, Dr. Shogo Kato from Tokyo Woman’s Christian University, Japan, and Dr. Yuuki Kato, Sagami Women’s University, Japan. This chapter focuses on the speed of interaction in text messaging via smart phone applications, specifically those with read receipt functions, and future directions in examining the strategic use of speed of interaction and read receipts. The next chapter, **“Narrowband Internet of Things,”** authored by Dr. Sudhir Routray of AASTU, Addis Ababa, Ethiopia, explores the main features of narrowband IoT (NB-IoT), the greener and more efficient version of IoT, and its potential and applications across different domains. The next chapter in this section, **“Application of Biogeography-based**

Optimization to Antennas and Wireless Communications,” authored by Prof. Sotirios Goudos of the Aristotle University of Thessaloniki, Greece, describes the biogeography-based optimization algorithm and presents its application to antenna and wireless communications design problems. The section ends with the chapter, **“Challenges and Opportunities of VLC Application in Intelligent Transportation Systems,”** authored by Profs. Aleksandra Kostic-Ljubisavljevic and Branka Mikavica of the University of Belgrade Serbia, which presents an analysis of the possibilities of VLC application in ITS scenarios and defines the main characteristics of VLC in ITS in terms of architecture, modulation, and standardization.

SECTION 8: DIGITAL MEDIA

This section features a collection of chapters that explores the effects of digital media as it becomes more engrained in today’s world and continues advancing rapidly along with its significant impact on individuals, society, and all facets of life on this earth. The first chapter in this section is, **“Antecedents and Consequences of Growing Popularity of Digital Consumption,”** authored by Dr. Ritu Narang and Miss Sonal Tiwari of the University of Lucknow, India. It proposes a conceptual model to extend knowledge about the antecedents and consequences of internet television consumption and its implications for individuals as well as society. The following chapter, **“Television Use and Consumption of Elderly Americans,”** authored by Profs. Robert Dunn and Stephen Marshall of East Tennessee State University, United States, presents research that offers a state of the art review on the elderly and their consumption of television, including new advancements in the media, potential motivations, and resulting effects. The next chapter, **“Social Media and Social Movements,”** authored by Ms. Sheldondra Brown and Dr. Monica Radu of Southeast Missouri State University, United States and Ms. Grace Babcock of North Carolina State University, United States, introduces the concepts of social media and social movements and also reviews recent literature examining how social media plays an active role in creating and encouraging those social movements. The following chapter, **“Digital Social Networks from a Social Capital Perspective,”** authored by Profs. Suparna Dhar and Mohammed Naved Khan of Aligarh Muslim University, India and Prof. Indranil Bose of the Indian Institute of Management, Kolkata, India, explores the unique opportunities offered by digitization of social networking, the novel capabilities of information dissemination and foraging on DSN sites, and the widespread adoption that necessitates theoretical interpretation of antecedents and consequences of DSN adoption. Concluding the section is the chapter, **“Exploring the Integration of User-Generated Content in Media Organizations Through Participatory Journalism,”** authored by Mrs. Theodora Saridou and Prof. Andreas Veglis of Aristotle University of Thessaloniki, Greece, which offers an in-depth description of the concept of participatory journalism, which holds an important and constantly evolving part in the digital media production.

SECTION 9: ENGINEERING SCIENCE AND TECHNOLOGIES

This section includes emerging research on technological tools and scientific methods that are expanding the field of engineering science. Opening this section is the chapter, **“Adapting a Requirements Engineering Process by Key Factors Estimation,”** authored by Dr. Graciela Hadad of both Universidad Nacional del Oeste and Universidad de Belgrano, Argentina, Prof. Jorge Doorn from Universidad Nacional del Oeste, Argentina, and Prof. Viviana Ledesma from Universidad Nacional de La Matanza, Argentina,

which proposes a process for constructing and dynamically adapting a requirements process, focusing on the evolving factors. The following chapter, **“Digital Tools Aimed to Represent Urban Survey,”** authored by Profs. Cristina Boido, Pia Davico, and Roberta Spallone of Politecnico di Torino, Italy, compares the knowledge and informative capabilities of different new technologies for urban modeling, through an overview of international researches, case studies, and personally conducted experiences. The next chapter in this section, **“Reengineering Legacy Systems Towards New Technologies,”** authored by Prof. Djelloul Bouchiha from Centre Universitaire De Naama SALHI Ahmed, Algeria, presents all the concepts, techniques and analyses in relation to the problem of re-engineering existing (legacy) systems toward new technologies. A following chapter, **“The Role of Two-Dimensional Materials in Electromagnetic Interference Shielding,”** authored by Dr. Rafael Vargas-Bernal from the Instituto Tecnológico Superior de Irapuato, Mexico, describes the impact that two-dimensional materials (or 2D materials) are having on the development of materials used for electromagnetic interference shielding, particularly the impulse of materials such as graphene, MXenes, transition metal dichalcogenides (TMDs), and phosphorene. The final chapter of the section, **“Modeling of the Aerodisperse Systems Hydrodynamics in Devices with Directional Motion of the Fluidized Bed,”** authored by Profs. Artem Artyukhov, Nadiia Artyukhova, and Ruslan Ostroha from Sumy State University, Ukraine and Prof. Jan Krmela from Alexander Dubcek University of Trencin, Slovakia, examines the usage of modern program packages for simulation of the aerodisperse systems hydrodynamics of equipment with directional motion of the fluidized bed for the chemical industry.

SECTION 10: ENVIRONMENTAL SCIENCE AND TECHNOLOGIES

This section highlights the newest innovations, strategies, and tools that are being used to mitigate pollution and examine the impacts of climate change in order to build a more sustainable global environment. The first chapter in this section, **“Renewable Resources and Value Based Complex Forest Management,”** authored by Dr. Yuri Pavlov from the Bulgarian Academy of Sciences, Bulgaria, investigates the question of multifactor value driven management of renewable resources of complex forest systems. A following chapter, **“Overview of Solid Waste Management of Healthcare and Related Organization,”** authored by Prof. Isaiah Oke from Obafemi Awolowo University, Nigeria, Prof. Lukman Salihu of University of Hafr Al-Batin, Saudi Arabia, Mr. Idi Mohammed from Land, Air and Water Consulting, Nigeria, and Prof. Asani Afolabi from Ladoke Akintola University, Nigeria, provides information on the quantities and properties of healthcare wastes in various types of facilities located in developing countries, as well as in some industrialized countries. The next chapter, **“Managing Water Resources by Industry Initiative,”** written by Dr. Sabyasachi Nayak of the Confederation of Indian Industry, India, seeks to capture the grassroots interventions undertaken by industry in improving the management of water resources. Another chapter in this section, **“Design and Implementation of a Location-Based Service with Emphasis on a Geographical Database,”** authored by Prof. Wen-Chen Hu of the University of North Dakota, United States, tries to mitigate the high difficulty of location-based service construction by showing the construction step by step with a focus on connecting a mobile device to a server-side geographical database. The final chapter in this section, **“Gendering Information and Communication Technologies in Climate Change,”** authored by Prof. Sam Wong of the University College Roosevelt, Netherlands, explains why, and how, women are more constrained than men from using ICTs in tackling climate change.

SECTION 11: ETHICS AND SOCIAL RESPONSIBILITY

This section provides thoughtful discussions on the ethical applications of technologies within society and the moral consequences and challenges they can present. The first chapter in this section is, **“Science and the Morality of Weapons Research,”** authored by Prof. John Forge of the University of Sydney, Australia. In this chapter the author attempts to answer the question: Is it ever morally justified to design the means to kill, harm and destroy, and if so, under precisely what circumstances? In a following chapter, **“Moral, Social and Political Responsibility in the Information Age,”** authored by Prof. Tomas Cahlik of Charles University Prague, Czech Republic, effects of ICTs on the responsibilities of living a good life and being a good person and having efficient and just social structures are discussed. Another chapter in this section, **“The Voice of Social Media, 1997-2018,”** authored by Dr. Bruce Cook of Cook Communications, United States, highlights a discussion on the consideration of social media censorship and questions of freedom of expression. In the next chapter, **“Breaking Fake News and Verifying Truth,”** authored by Dr. Kazuhiko Shibuya of ROIS, Japan, online fact-checking studies on fake news are reviewed. The concluding chapter in this section, **“Politics and Ethics and a National Framework to Combat Corruption in Zimbabwe,”** authored by Prof. Tawanda Zinyama from the University of Zimbabwe, Zimbabwe, seeks to interrogate issues of politics and ethics because they have a bearing on the realization or lack thereof of the Zimbabwean Vision 2030.

SECTION 12: GOVERNMENT, POLICY, AND CRISIS MANAGEMENT

This section examines government policies and practices and explores future opportunities and lessons learned from recent disasters around the world for the development of better crisis response management. Opening this section is the chapter, **“Good Practices in e-Government Accessibility,”** authored by Profs. Fernando Almeida and José Monteiro from the Polytechnic Institute of Gaya, Portugal, which looks to the accessibility of good practices that should be adopted in the development of e-Government platforms. The next chapter in this section, **“National Systems of Innovation Using an Application on EU Data,”** authored by Prof. George Korres from the University of Aegean, Gabon and Prof. Maria Michailidis from the University of Nicosia, Cyprus, reviews the theory and the current literature of the national systems of innovation and attempts to apply some statistical measurement and indices in order to estimate the effects and the implications on EU innovation systems member states. The following chapter, **“Comparative Analysis of Applying Behavioral Public Policy to Telecommunication Market by International Organizations,”** authored by Dr. Nagayuki Saito from the International Professional University of Technology in Tokyo, Japan, reviews the status of behavioral public policy being introduced in each international organization and compares and verifies the stages of behavioral public policy in terms of consumer protection policy in each international organization in the telecommunications market. The next chapter, **“A Good American President,”** authored by Mr. James George, Archway Publishing, United States and Mr. Abdullah Murrar, Dr. Pankaj Chaudhary, and Dr. James Rodger from the Indiana University of Pennsylvania, United States, describes political engineering at an atomic level; that is, addressing what makes a competent elected official and using the office of president of the United States as an example. The concluding chapter of this section, **“A Risk Management on Demographic Mobility of Evacuees in Disaster,”** authored by Dr. Kazuhiko Shibuya of ROIS, Japan, describes both

risks and crises on disaster and specifically reviews issues and discusses future design for the lessons from the Fukushima nuclear power plant accident.

SECTION 13: HUMAN ASPECTS OF TECHNOLOGY

This section discusses societal utilization of technology and the impact it has on social and economic development across the globe. Opening this section is the chapter, **“The Technology and the Theory of Apocalypse,”** authored by Prof. Maximiliano Korstanje of the University of Palermo, Argentina. It ignites an interesting discussion around the idea of apocalypse connection to the role of technology in slicing humanity from nature. The next chapter in this section, **“Software Literacy as a Vital Digital Literacy in a Software-Saturated World,”** authored by Prof. Craig Hight from the University of Newcastle, Australia and Prof. Elaine Khoo of The University of Waikato, New Zealand, proposes the notion of ‘software literacy’ to highlight a neglected aspect of digital literacies and a key means of conceptualizing the skills and understandings needed for people to be critical and creative users of software in today’s software saturated culture. The following chapter, **“ICT Investments on Economic Sectors with International Comparative Advantage and the Diffusion of Prosperity,”** authored by Dr. Apostolos Syropoulos of the Greek Molecular Computing Group, Greece, shows how Information Communication and Technologies (ICTs) can be used to boost the economy of a country that emerges from a deep crises. In the next chapter, **“Dimensions of the Digital Divide,”** authored by Prof. Marcus Leaning from the University of Winchester, United Kingdom and Mr. Udo Averweg of eThekweni Municipality, South Africa, the authors claim that the idea of the digital divide can be understood to operate in a range of different ‘dimensions’ and argue that the problem is a multi-faceted one which requires a multi-level approach to address. The concluding chapter in this section, **“How Mobile Technologies are Leading to Economic Development in Sub-Saharan Africa,”** authored by Prof. Nigel McKelvey and Mr. Adam Crossan from the Letterkenny Institute of Technology, Ireland and Prof. Kevin Curran from Ulster University, United Kingdom, investigates the use of mobile technology to help residents of sub-Saharan Africa in the sectors of health care and education.

SECTION 14: LIBRARY SCIENCE AND RESEARCH METHODS

This section presents the latest research findings in methodologies and techniques in managing informational competencies and Q-method, as well as the impact that technology has had on libraries worldwide. The first chapter in this section, **“Informational Competencies,”** authored by Prof. José Rascão from the Polytechnic Institute of Setúbal, Portugal, studies the informational competences and processes of its construction, transmission, and use, based on the aggregation of existing literature. The following chapter, **“A primer on Q-method and the Study of Technology,”** authored by Dr. Stéphanie Gauttier of Grenoble Ecole de Management, France, provides an in-depth description of Q-method and demonstrates how its use could be beneficial to studies of technology and could reinforce the transparency and validity of other qualitative methods. The next chapter in this section, **“Modernizing the Academic Library,”** authored by Prof. Jennifer Joe of University of Toledo, United States, examines libraries and their response to change and thinking outside of the box when it comes to their employees, their resources, and their mission statements in order to meet the challenges that the digital age presents. The

following chapter in this section, **“Prospects and Challenges of Web 3.0 Technologies Application in the Provision of Library Services,”** authored by Prof. Promise Ilo, Mr. Roland Izuagbe, and Mr. Michael Fagbohun, from Covenant University, Nigeria, Prof. Christopher Nkiko of Elizade University, Nigeria, and Profs. Cyprian Ugwu and Ngozi Ekere from the University of Nigeria Nsukka, Nigeria, provides an in-depth description of Q-method and demonstrates how its use could be beneficial to studies of technology and could reinforce the transparency and validity of other qualitative methods. The final chapter in this section, **“Massive Digital Libraries (MDLs) and the Impact of Mass-Digitized Book Collections,”** authored by Mr. Andrew Weiss of California State University, Northridge, United States, describes the characteristics of Massive Digital Libraries (MDLs) and outlines their impact upon current information science issues, especially digital collection metadata, copyright and fair use, the diversity of source collections, and user privacy.

SECTION 15: MEDICAL ETHICS AND LAW

This section contains forward-thinking research on new medical policies and laws as well as ethical issues that have recently been presented in the regulation of medical devices, technologies, and patient information. The opening chapter of this section is, **“Protocol Development in Clinical Trials for Healthcare Management,”** authored by Dr. Swati Jagdale and Dr. Anuruddha R. Chabukswar from the Dr. Vishwanath Karad MIT World Peace University, India and Miss Asawaree Anand Hable of MAEER’s Maharashtra Institute of Pharmacy, India, which discusses the development of protocols for clinical trials to serve as a roadmap for healthcare professionals involved in investigation work. The following chapter, **“Ambiguities in the Privacy Policies of Common Health and Fitness Apps,”** authored by Prof. Devjani Sen from Algonquin College, Canada and Prof. Rukhsana Ahmed from the University of Albany, SUNY, United States, provides a set of recommendations when designing privacy policies for the sharing of personal information collected from health and fitness apps. The next chapter, **“Generosity in Healthcare Policy under the Obama Administration,”** authored by Prof. Khadijeh Rouzbehani from the University of Victoria, Canada, examines the Affordable Care Act as an improved alternative to the previous policy regime on healthcare in the United States. The following chapter, **“Islamic Medical Ethics in Brunei Darussalam Hospitals,”** authored by Prof. Syazana Fauzi of Universiti Brunei Darussalam, Brunei, seeks to analyze whether Brunei Darussalam’s health professionals handle medical ethical cases in a manner that is congruent to the State Mufti’s fatāwā. The final chapter in this section, **“Distribution of Selected Health Technology in Regions of Slovakia,”** authored by Profs. Beata Gavurova and Matus Kubak from the Technical University of Košice, Slovakia, deals with the issue of regional disparities in the use of selected medical technology.

SECTION 16: MEDICAL TECHNOLOGIES

This section contains the most recent research on the latest technologies utilized and applied in healthcare for the more efficient and effective diagnosis and treatment of diseases and illnesses. The first chapter in this section, **“Expanding Scope of Information Technology in Clinical Care,”** authored by Dr. Gumpeny Sridhar of the Endocrine and Diabetes Centre, India, describes the genesis and evolution of an electronic medical record (EMR) system at the Endocrine and Diabetes Centre and proposes ways

in which the existing system can be further developed with additional features. A following chapter, **“Artificial Neural Networks in Medicine - Recent Advances,”** authored by Prof. Steven Walczak from the University of South Florida, United States, examines recent trends and advances in Artificial Neural Networks (ANN) and provides references to a large portion of recent research, as well as looking at the future direction of research for ANN in medicine. The next chapter, **“Internet of Things and Data Science in Health Care,”** authored by Dr. George Tzanis and Mrs. Ourania-Ioanna Fotopoulou of Aristotle University of Thessaloniki, Greece, introduces the Cloud-IoT paradigm in the healthcare domain in order to offer manifold benefits and opportunities that will considerably improve the quality of health care. The concluding chapter, **“Highly Aged and After Cloud,”** authored by Dr. Shigeki Sugiyama an Independent Researcher from Japan, presents research on the latest technological solutions in managing the most effective and proper treatment of highly aged people in Japan. The final chapter of this section, **“Gene Editing Technology and Ethical Issues,”** authored by Ms. Barbara Holland from the Brooklyn Public Library, United States, reviews gene editing technology and its study and treatment of human diseases, as well as its ethical concerns.

In summary, the primary research contents included in this reference publication is aimed to provide a much better understanding of the issues, trends, and solutions surrounding the themes of its sixteen sections, providing the coverage of the most recent research conducted by many researchers from all over the world. Collectively, the authors of the contents of this encyclopedia bring about the most current thoughtful and comprehensive overview of many important topics related to the field of information science and technology and its continuous impact on all facets of the world we are living in. We hope that their diverse and comprehensive coverage offered within their chapters will contribute to a broader awareness of all topics, research, discoveries, and solutions related to information science and technology. We are also hopeful that the contents of this publication will inspire other researchers to further explore the expansion of existing research to discover new boundaries and opportunities in the field that can be further developed to better serve the inhabitants of our planet.

Mehdi Khosrow-Pour, D.B.A.

Editor-in-Chief

Encyclopedia of Information Science and Technology, Fifth Edition

Acknowledgment

Editing and completing an authoritative and comprehensive scholarly research publication such as the *Encyclopedia of Information Science and Technology, Fifth Edition* requires tremendous contributions and a great deal of assistance from large groups of scholars and staff. The primary objective of this encyclopedia is to provide the most up-to-date scholarly coverage of all topics related to the evolving nature of information science and technology, and to carry on the tradition of the four editions that successfully preceded this one.

The contributed chapters from expert researchers from all over the world provide an in-depth look into topics such as computational science, digital terrorism, cyber security and data protection, digital and wireless communications, environmental science and technologies, ethics and social responsibility, human aspects of technology, library science, medical technologies, and much more. I am indebted to all the authors for their excellent contributions to this edition of the *Encyclopedia of Information Science and Technology, Fifth Edition*.

All submitted chapter manuscripts to this publication underwent a double-blind peer review process in order to achieve the highest level of validity, reliability, accuracy, and quality of the submitted research. I am thankful to all the reviewers of this encyclopedia for providing their expertise and their rigorous, unbiased assessment of the chapter manuscripts assigned to them on a double-blind basis, as well as the members of the Editorial Advisory Board for their wisdom, guidance, and assistance with various decisions throughout the editorial process.

I would also like to convey my deepest appreciation and gratitude to all those individuals who assisted me in editing this publication, which include Lindsay Wertman, Managing Director of IGI Global, Jan Travers, Director of Intellectual Property and Contracts at IGI Global, Melissa Wagner, Managing Editor of Acquisitions at IGI Global, Maria Rohde, Assistant Development Editor at IGI Global, Crystal Moyer, Assistant Development Editor at IGI Global, Eric Whalen, Assistant Development Editor at IGI Global, Josh Christ, Assistant Development Editor at IGI Global, Chris Shearer, Copy Editing Manager at IGI Global, and Michael Brehm, Managing Editor of Book and Journal Production at IGI Global. Additionally, I would like to thank the IGI Global Sales and Marketing Department, especially Nick Newcomer, Senior Director of Marketing and Sales at IGI Global, Caroline Campbell, Marketing Manager at IGI Global, Brittany Haynes, Marketing Editorial Assistant at IGI Global, and Tori Parks, Graphic Design Coordinator at IGI Global, for their endless support in promoting this invaluable reference source.

Thanks to everyone who has provided me immeasurable amounts of knowledge, wisdom, and support over the last 30 years.

Acknowledgment

Mehdi Khosrow-Pour, D.B.A.

Editor-in-Chief

Encyclopedia of Information Science and Technology, Fifth Edition

Section 1

Artificial Intelligence

Autonomous Vehicles

1

Yair Wiseman

 <https://orcid.org/0000-0002-4221-1549>

Bar-Ilan University, Israel

INTRODUCTION

The first car was invented in 1870 by Siegfried Marcus (Guarnieri, 2011). Actually it was just a wagon with an engine but without a steering wheel and without brakes. Instead, it was controlled by the legs of the driver.

Converting traditional vehicles into autonomous vehicles was not just one step. The first step was just 28 years after the invention of cars that is to say 1898. This step's concept was moving a vehicle by a remote controller (Nikola, 1898). Since this first step and as computers have been becoming advanced and sophisticated, many functions of modern vehicles have been converted to be entirely automatic with no need of even remote controlling.

Changing gears was one of the first actions that could be done automatically without an involvement of the driver (Anthony, 1908), so such cars got the title of “automatic cars”; however, nowadays there are vehicles that can completely travel by themselves although they are not yet allowed to travel on public roads in most of the world. Such vehicles are called “autonomous vehicles” or “driverless cars”.

BACKGROUND

Autonomous vehicles will noticeably change the worldwide transport market. Autonomous vehicles will improve our quality of life and road safety. The number of traffic accidents will be significantly reduced. In Addition, autonomous vehicles can improve on the parking issues like the ability to park in a remote location in a rural fringe of the city. These vehicles will be more fuel efficient and their insurance will be cheaper. Also, people with disabilities can greatly benefit from this technology and military autonomous vehicles can prevent injuries and deaths in combat.

There are many practical obstacles and law difficulties that should be discussed. There were several car accidents where autonomous vehicles were involved (Wiseman & Grinberg, 2016a), (Wiseman & Grinberg, 2016b), (Wiseman & Grinberg, 2016c), (Wiseman & Grinberg, 2018). Until now two of these accidents were very tragic resulting in the death of a person.

There are several companies along with several government branches (especially military branches) that develop autonomous vehicles. Many of their products will be available for use in the coming years.

There is a need for more research until a car will be able to safely take us to our destination, without being involved in an accident and without having to drive or be there at all. Where is our car on the scale of autonomy and advanced safety? The international automotive organization SAE International (Society of Automotive Engineers) has defined six different levels of autonomous driving, or independent, in vehicles (SAE, 2014). This scale is quite acceptable in the autonomous vehicle industry. These six levels are divided into two groups: The three low levels require the driver to be active at least in part. In

DOI: 10.4018/978-1-7998-3479-3.ch001

the high three levels, there may be situations where the driver is not required to drive at all. These six levels are listed herein below:

Level 0: No Driving Automation. The driver is in charge of all driving activities in other words a vehicle which is not autonomic at all.

Level 1: Driver Assistance. One autonomic system is operating in the vehicle. For example, Electronic Stability Control (ESC) (Lie, Tingvall, Krafft & Kullgren, 2006), Adaptive Cruise Control (ACC) (Van Arem, Van Driel & Visser, 2006), Emergency Brake Assist (EBA) (Page, Foret-Bruno & Cuny, 2005). The autonomic system acts by itself as a help for the driver. However, the autonomic system is not used for driving, but only helps the driver when necessary, as an emergency device.

Level 2: Partial Driving Automation. A vehicle in which two or more autonomic systems work together, capable of simultaneously handling two tasks. For example both steers the vehicle and operate the braking system. The main limitation of this level is that in any given moment, the driver must analyze the conditions of the road and respond accordingly. Thus, even if the car is of level 2, the technology still needs the driver to be aware of the circumstances and to respond immediately to any scenario because the autonomic system cannot prevent some of the potential accidents.

Level 3: Conditional Driving Automation. In some conditions the driver is not required to monitor the driving; however, there are limitations in this level; in other words the car cannot be driven without human intervention under any conditions. For example it can be automatically driven only on highways. Accordingly, when the vehicle goes, a human driver must supervisor it, because if emergency circumstances come about, the human driver will intervene so as to avoid any potential accident if the car cannot handle the circumstances.

Level 4: High Driving Automation. The vehicle will be able under certain circumstances to operate fully autonomously and to protect its passengers from accidents. An autonomous vehicle at this level does not require a human driver to be ready to respond in an emergency, but only if the defined circumstances are kept.

Level 5: Full Driving Automation. There is no restriction to certain circumstances, nor is there a need for a human driver for supervision in emergency circumstances.

MOTIVATION

Approximately 94-95 percent of road accidents are caused by human errors (NHTSA, 2015), (Freeman, 2016). These errors are usually taken place because of smartphone use, fatigue or distraction because of any other reason (Kaplan, Guvensan, Yavuz & Karalurt, 2015). An autonomous vehicle will prevent the vast majority of them. In addition, an autonomous vehicle will always conduct itself as it has been programmed. Therefore, if a programming failure is detected that has not been occurred in all tests, it will be possible to issue an update how to correct this failure. This update will instruct all autonomous cars how to conduct themselves in a similar case (Koopman & Wagner, 2016). In this way, every autonomous car will be trained by the experience of all autonomous cars as opposed to a human driver trained only by his own experience.

Autonomous cars will be of great help for people with disabilities (Bradshaw-Martin, & Easton, 2014). For example, people with severe visual impairments can safely travel in an autonomous car. In addition, people with an exhausting job or drunk people will be able to travel in such a car without driving. Moreover, autonomous vehicles will release also people in good physical shape from the burden

of driving and will let them do other activities during their travels like work, rest or just enjoy the view (Litman, 2017).

Another issue is air pollution which will be moderated as well. Such smart junctions, whose traffic is routed on the basis of allocation of slots to pass the junction, will significantly reduce traffic jams and other delays (Fayazi, 2017). Such junctions are expected to have an important impact on the road system of each city. The travel and waiting times will be considerably shorter (Zohdy, Kamalanathsharma & Rakha, 2012). Fuel consumption will be decreased and the traffic will flow more freely, with less stoppings and accelerations, which will significantly reduce air pollution (Talebpour & Mahmassani, 2016). In addition, autonomous vehicles will be driven in a more conservative manner; therefore autonomous vehicles will consume substantially less gas and energy while travelling, compared to a traditional vehicle driven by a human, because a substantial amount of gas is consumed when driving at speed that is too high for the vehicle, accelerating excessively or unnecessary braking. Autonomous vehicles will eliminate these undesirable ways of driving; meaning less gas is consumed resulting in less air pollution.

Autonomous vehicles will put an end to the bother of hectic parking systems (Wiseman, 2017a), (Wiseman, 2017b). Furthermore, autonomous vehicles will change the land uses in the metropolitan areas anywhere (Brett, 2016). Many large parking lots will be shifted into alternative uses. (Shoup, 1997) argues that nearly a third of the central area in major cities in the US is reserved for parking. Autonomous vehicles can improve on the parking problems because of several main value-added aspects:

- Unlike traditional vehicles, autonomous vehicle will be able to drop off passengers in almost any spot in any city and then instead of cruising for a free parking (Inci, 2015), the autonomous vehicle will park itself in a remote parking lot in a rural fringe of the city where low-cost land is more readily available (Tayade & Patil, 2016); therefore there will be no necessity, for instance, for an enormous parking lot near airports as opposed to the situation that exists nowadays.
- An autonomous vehicle can park itself in a tighter fitting parking slot without scrapping nearby vehicles (Grinberg & Wiseman, 2013), (Grinberg & Wiseman, 2007). In (Bertoncello & Wee, 2015), it has been estimated that the size of a typical parking lot can be reduced by 15% because of the tighter fitting parking slots of autonomous vehicles. Therefore, autonomous cars will enable to build the parking lots in a more compact manner.
- An autonomous vehicle will be parked by the computer software (Filatov & Serykh, 2016); therefore, there will be no more occurrences of drivers parking their vehicles in an angle in a perpendicular parking; nor will there be drivers parking their cars in other ways that occupy more than one parking slot. In addition, the parking slots will be allocated according to the vehicle size and as a result the parking slots will not be any longer at the size of the largest potential vehicle (Li & Shao, 2015).

SOLUTIONS AND RECOMMENDATIONS

Traffic lights were invented at 1868 (Wang, 2016) and they have been turning out to be more sophisticated and computer controlled (Fang, 2016). The new generation of traffic lights is called “intelligent traffic lights” and it contains digital electronic units derived from common computers (Wiseman, 2016a), (Wiseman, 2016b). The traffic lights are a common sight in the streets of our cities. They use up space in the streets, need a maintenance team and consume electricity power; however, the traffic lights will probably disappear. Autonomous vehicles equipped with sensors will be able to communicate with each

other (Wiseman, Schwan & Widener, 2005). Moreover, they will be able to maintain a safe distance between them with the purpose of going across junctions without using traffic lights. Similarly to the model of air traffic control systems, computerized junctions can route traffic on the basis of time slot allocation, as a result the traffic lights can be replaced by a computerized communication system (Tachet, 2016). When an autonomous vehicle approaches a junction, an access request to the junction will automatically be sent from the autonomous vehicle to a traffic management system. This system will receive the request and assign it a unique time slot to pass the junction.

Likewise, Autonomous vehicles will be able to coordinate their way of driving, so they can move in platoon with fixed intervals (Tsugawa, Kato, & Aoki, 2011), by employing Lidar and ultrasonic range-finders (Wiseman, 2018) with the intention of distance reduction between them, without any risk of collisions. These platoons will reduce traffic congestion on roads, will increase the energy efficiency of driving and will reduce the air pollution generated by vehicles (Fernandes & Nunes, 2012).

DIFFICULTIES AND OBSTRUCTIONS

The Luddites (Jones, 2013) were a movement of English workers in the beginning of the 19th century who protested, sometimes even by violence measures like destroying machines, against the changes brought about by the Industrial Revolution, which they saw as a threat to their livelihood. Similarly, Autonomous vehicles can cause mass joblessness in the automotive industry. Professional drivers like taxi drivers or truck drivers will not be required anymore. Likewise the Luddites, some unhappy professional drivers attacked autonomous vehicles in California (Wong, 2018). This joblessness will also occur in the garage industry. When 95 percent of the vehicular accidents are eliminated, a much small number of garages will be required and a large amount of automobile mechanics will be fired and will not be able to find a new job. Significantly lower rate of vehicular accidents will also cause less damage to road infrastructure, bringing about some reduction in need for workers in the road maintenance sector.

Another issue that can be challenging is cyber criminalism or cyber terrorism (Petit, & Shladover, 2015). In the future, almost all if not entirely all the vehicles will be exclusively controlled by computers. If the software is not perfect and some flaws make the software vulnerable, smart hackers might be capable to unlawfully interfere with the vehicle software, so they will be able to harm the vehicle or even worse harm the vehicle occupants. In addition, a new version of car thefts may possibly be a fleet of autonomous vehicles thefts.

The trouble of traffic congestion is very acute in many cities and highways (Wiseman, 2017c), (Wiseman, 2017d), (Wiseman, 2017e). The autonomous vehicles will add many cars to the already congested roads, because of several reasons:

- Children are a significant portion of the world population; however, they mostly do not own a vehicle; nor do they drive a vehicle. When autonomous cars are available, the children will be able to own a vehicle and to “drive” vehicles. It is hard to predict which portion of the children population will use private cars instead of the public transportation (Fagnant & Kockelman, 2015), but even if a small percentage of the children are added to the already congested roads, it might augment the congestion difficulty.
- Similarly to the children, people with a disability like blindness or other severe disability that nowadays prevents them from driving a vehicle, will be able to use an autonomous vehicle by themselves. Actually, this happening was one of the first motivations for the developing of autono-

mous cars because such a car will definitely and substantially improves on the disabled quality of life (Madarasz, Heiny, Crompt & Mazur, 1986); however, this improvement will add more vehicles on roads.

- Some people prefer to travel by public transportation because they like to relax while going to their destination. Some of them want to sleep, read or just not arriving in an edgy mood after conflicts with other drivers while driving. The autonomous vehicle will give these people the opportunity to have this rest with no need of public transportation. Traveling by an autonomous vehicle will be usually even a better relaxation because there will be no need for connections i.e. to change from bus to train etc. nor will there be other people in the vehicle. This will make more people to choose the autonomous vehicles.
- More traffic will be added because empty vehicles will go alone in their way to park. Nowadays, drivers make an effort to find a close parking lot, so they will not have to walk a large distance; however, an autonomous vehicle can go by itself to a remote parking lot.

In spite of this added traffic, autonomous vehicles will drive more efficiently, so there are more than a few analyses which have confidence that this efficiently way of driving can counterbalance the added traffic and even obtaining a better traffic flow.

FUTURE RESEARCH DIRECTIONS

The current world population is more than 7.6 billion; however it is expected to reach 10 billion in the year 2056 (Worldometers, 2018). In addition, the standard of living has been increasing in most of the world countries and it is expected that more people will be able to own a car in the coming centuries (Reich, 2010). The reasons mentioned in the previous section for potential excessive congestions when autonomous cars are available will also play a part in the even now congested roads, that is to say many current roads are already congested and even if a major road construction is performed and even a second floor is built in the most congested roads, the problem of traffic congestion will be acute. This concern of congested roads has already stimulated enterprises all over the world to make the first moves toward flying cars.

A flying car will be a private air vehicle which can travel both by air and roads. Many prototypes of such cars like AeroMobil (Becker, 2017), Urban Aeronautics' XHawk (Yoeli, 2002), The Skycar (Pecora, Pecora & Lecce, 2011), The Xplorair (Altran France, 2013) and many more (Frost & Sullivan, 2017) have been built and have been tested; however, the production of these flying cars has not yet started for a widespread use.

The concept is a vehicle that is able to fly and to vertical take-off and land (VTOL). Such a vehicle will be able to go without a road and also several levels of elevation can be attained, as a result there will be a multiple level road without a need to pave even a single mile of road.

Autonomous flying cars will also have the ability to connect and to move in platoons, so they will be able to catch sight of a much larger area, so they will be much safer and more efficient (Wang, Xu, Leng, & Pollin, 2018, June).

The overall consequences of autonomous vehicles are still unknown and the impact on other means of transportation is unclear (Wiseman, 2019a). Also, the impact on transportation infrastructure is undecided (Wiseman, 2019b). When the autonomous vehicles penetrate the transportation market, more information will be available and this information will enable more profound research.

CONCLUSION

Autonomous vehicles are self-guided vehicles that overcome obstacles without any help of a human driver. These vehicles are expected to go on the same road and with the same traffic signs we are currently familiar with. Autonomous vehicles are equipped with several apparatus that enable them to observe their surroundings and decide in real time what action should be taken.

Autonomous vehicles are expected to be safer than vehicle driven by human drivers. Furthermore, autonomous vehicles will release people from the burden of driving and will give them more time to rest, work or do other activities. These vehicles can be operated by children or other incapable persons. In addition, they will be driven in more efficient manner than an average driver. Parking concerns will be enormously affected by autonomous vehicles because autonomous vehicles can park themselves in a remote location and in a tighter fitting parking slot.

There are some concerns like many professional drivers who will lose their jobs and the potential congestions ensued by additional vehicles going empty or going with children; however with a proper preparations the shift to autonomous vehicles can be done smoothly and efficiently.

REFERENCES

- Altran France. (2013). *Xplorair: un projet Altran Research - Altran France*. Available online at: https://www.youtube.com/watch?v=_v4FNXvF9kQ
- Anthony, J. S. (1908). *Automatic speed-changing gear*. U.S. Patent No. 907,711. Washington, DC: U.S. Patent and Trademark Office.
- Becker, E. P. (2017). The future of flying is near. *Tribology & Lubrication Technology*, 73(8), 96.
- Bertoncello, M., & Wee, D. (2015). Ten ways autonomous driving could redefine the automotive world. McKinsey & Company.
- Bradshaw-Martin, H., & Easton, C. (2014). Autonomous or 'driverless' cars and disability: a legal and ethical analysis. *European Journal of Current Legal Issues*, 20(3).
- Brett, J. A. (2016). *Thinking local about self-driving cars: a local framework for autonomous vehicle development in the united states* (MSc Thesis). Urban Design and Planning, University of Washington.
- Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A, Policy and Practice*, 77, 167–181. doi:10.1016/j.tra.2015.04.003
- Fang, J. (2016). Intelligent traffic light controller design using FPGA digest of technical papers. In *Proceedings of IEEE International Conference on Consumer Electronics (ICCE 2016)*, (pp. 449-452). IEEE.
- Fayazi, S. A., Vahidi, A., & Luckow, A. (2017, May). Optimal scheduling of autonomous vehicle arrivals at intelligent intersections via MILP. In *Proceedings of American Control Conference (ACC 2017)*, (pp. 4920-4925). 10.23919/ACC.2017.7963717
- Fernandes, P., & Nunes, U. (2012). Platooning with IVC-enabled autonomous vehicles: Strategies to mitigate communication delays, improve safety and traffic flow. *IEEE Transactions on Intelligent Transportation Systems*, 13(1), 91–106. doi:10.1109/TITS.2011.2179936

Filatov, D. M., & Serykh, E. V. (2016, May). Intelligence autonomous parking control system of four-wheeled vehicle. In *Proceedings of IEEE International Conference on Soft Computing and Measurements (SCM 2016)*, (pp. 154–156). 10.1109/SCM.2016.7519713

Flinn, E. D. (2003). Revolutionary X-Hawk hovers near success. *Aerospace America*, 41(7), 26–28.

Freeman, D. (2016). Self-Driving Cars Could Save Millions Of Lives — But There's A Catch. *HuffPost*. Available online at: https://www.huffingtonpost.com/entry/the-moral-imperative-thats-driving-the-robot-revolution_us_56c22168e4b0c3c550521f64

Frost & Sullivan. (2017). *Future of Flying Cars, 2017–2035*. Available online at: <http://www.frost.com/sublib/frost-content.do?sheetName=report-overview&sheetGroup=K16A-01-00-00-00&viewName=virtual-brochure&repid=K16A-01-00-00-00>

Grinberg, I., & Wiseman, Y. (2007, August). Scalable parallel collision detection simulation. *Proceedings of Signal and Image Processing*, 7, 380–385.

Grinberg, I., & Wiseman, Y. (2013). Scalable parallel simulator for vehicular collision detection. *International Journal of Vehicle Systems Modelling and Testing. Inderscience Publications*, 8(2), 119–144.

Guarnieri, M. (2011). When Cars Went Electric, Part 2. *IEEE Industrial Electronics Magazine*, 5(2), 46–53. doi:10.1109/MIE.2011.941122

Inci, E. (2015). A review of the economics of parking. *Economics of Transportation*, 4(1-2), 50–63. doi:10.1016/j.ecotra.2014.11.001

Jones, S. E. (2013). *Against technology: From the Luddites to neo-Luddism*. Routledge.

Kaplan, S., Guvensan, M. A., Yavuz, A. G., & Karalurt, Y. (2015). Driver behavior analysis for safe driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 16(6), 3017–3032. doi:10.1109/TITS.2015.2462084

Koopman, P., & Wagner, M. (2016). Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 4(1), 15–24. doi:10.4271/2016-01-0128

Li, B., & Shao, Z. (2015). A unified motion planning method for parking an autonomous vehicle in the presence of irregularly placed obstacles. *Knowledge-Based Systems*, 86, 11–20. doi:10.1016/j.knysys.2015.04.016

Lie, A., Tingvall, C., Krafft, M., & Kullgren, A. (2006). The effectiveness of electronic stability control (ESC) in reducing real life crashes and injuries. *Traffic Injury Prevention*, 7(1), 38–43. doi:10.1080/15389580500346838 PMID:16484031

Litman, T. (2017). *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute.

Madarasz, R., Heiny, L., Crompt, R., & Mazur, N. (1986). The design of an autonomous vehicle for the disabled. *IEEE Journal on Robotics and Automation*, 2(3), 117–126. doi:10.1109/JRA.1986.1087052

NHTSA - National Highway Traffic Safety Administration. (2015). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. NHTSA's National Center for Statistics and Analysis, United States Department of Transportation, DOT HS 812 115. Available online at: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115>

- Nikola, T. (1898). *U.S. Patent No. 613,809*. Washington, DC: U.S. Patent and Trademark Office.
- Page, Y., Foret-Bruno, J. Y., & Cuny, S. (2005). Are expected and observed effectiveness of emergency brake assist in preventing road injury accidents consistent? *Proceedings of the 19th International Technical Conference on the Enhanced Safety of Vehicles (ESV Conference)*.
- Pecora, R., Pecora, M., & Lecce, L. (2011, September). Flutter certification of SKYCAR aircraft: Rational analysis and flight tests. *Proceedings of the 3rd CEAS Air&Space Conference*.
- Petit, J., & Shladover, S. E. (2015). Potential cyberattacks on automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 16(2), 546–556.
- Reich, R. B. (2010). *The Work of Nations: Preparing Ourselves for 21st Century Capitalis*. Vintage Books.
- SAE. (2014). Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems-J3016. In Society of Automotive Engineers: On-Road Automated Vehicle Standards Committee. SAE (Society of Automotive Engineers) Pub. Inc.
- Shoup, D. C. (1997). The high cost of free parking. *Journal of Planning Education and Research*, 17(1), 3–20. doi:10.1177/0739456X9701700102
- Tachet, R., Santi, P., Sobolevsky, S., Reyes-Castro, L. I., Frazzoli, E., Helbing, D., & Ratti, C. (2016). Revisiting street intersections using slot-based systems. *PLoS One*, 11(3), e0149607. doi:10.1371/journal.pone.0149607 PMID:26982532
- Talebpour, A., & Mahmassani, H. S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C, Emerging Technologies*, 71, 143–163. doi:10.1016/j.trc.2016.07.007
- Tayade, Y., & Patil, M. D. (2016). Advance Prediction of Parking Space Availability and Other Facilities for Car Parks in Smart Cities. *International Research Journal of Engineering and Technology*, 3(5), 2225–2228.
- Tsugawa, S., Kato, S., & Aoki, K. (2011). An automated truck platoon for energy saving. *Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4109–4114. 10.1109/IROS.2011.6094549
- Van Arem, B., Van Driel, C. J., & Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Transactions on Intelligent Transportation Systems*, 7(4), 429–436. doi:10.1109/TITS.2006.884615
- Wang, M. Y., Ding, H. W., Zhao, Y. F., Liang, Z. G., & Zhu, L. Z. (2016). The Application of the Exhaustive Polling Theory in Intelligent Traffic System. In *MATEC Web of Conferences (Vol. 44, paper 02013)*. EDP Sciences. 10.1051/mateconf/20164402013
- Wang, Q., Xu, C., Leng, S., & Pollin, S. (2018, June). When Autonomous Drones Meet Driverless Cars. *Proceedings of the 16th ACM Annual International Conference on Mobile Systems, Applications, and Services*, 514.
- Wiseman, Y. (2016). Conceptual Design of Intelligent Traffic Light Controller. *International Journal of Control and Automation*, 9(7), 251–262. doi:10.14257/ijca.2016.9.7.23

- Wiseman, Y. (2016). Traffic Light with Inductive Detector Loops and Diverse Time Periods. *Contemporary Research Trend of IT Convergence Technology*, 4, 166–170. doi:10.21742/asehl.2016.4.39
- Wiseman, Y. (2017). Remote Parking for Autonomous Vehicles. *International Journal of Hybrid Information Technology*, 10(1), 313–324. doi:10.14257/ijhit.2017.10.1.27
- Wiseman, Y. (2017). Self-Driving Car-A Computer will Park for You. *International Journal of Engineering & Technology for Automobile Security*, 1(1), 9–16. doi:10.21742/ijetas.2017.1.1.02
- Wiseman, Y. (2017). Real-time monitoring of traffic congestions. *IEEE International Conference on Electro Information Technology (EIT 2017)*, 501–505. 10.1109/EIT.2017.8053413
- Wiseman, Y. (2017). Tool for Online Observing of Traffic Congestions. *International Journal of Control and Automation*, 10(6), 27–34. doi:10.14257/ijca.2017.10.6.04
- Wiseman, Y. (2017). Computerized Traffic Congestion Detection System. *International Journal of Transportation and Logistics Management*, 1(1), 1–8. doi:10.21742/ijtlm.2017.1.1.01
- Wiseman, Y. (2018). Ancillary Ultrasonic Rangefinder for Autonomous Vehicles. *International Journal of Security and its Applications*, 12(5), 49–58.
- Wiseman, Y. (2019). Driverless Cars will Make Passenger Rails Obsolete, IEEE. *Technology in Society*, 38(2), 22–27.
- Wiseman, Y. (2019). Driverless Cars will Make Union Stations Obsolete. *The Open Transplantation Journal*, 13(1), 109–115. doi:10.2174/1874447801913010109
- Wiseman, Y., & Grinberg, I. (2016, May). Circumspectly crash of autonomous vehicles. *IEEE International Conference on Electro Information Technology (EIT 2016)*, 387–392. 10.1109/EIT.2016.7535271
- Wiseman, Y., & Grinberg, I. (2016). Autonomous vehicles should not collide carelessly. *Advanced Science and Technology Letters*, 133, 223–228.
- Wiseman, Y., & Grinberg, I. (2016). When an inescapable accident of autonomous vehicles is looming. *International Journal of Control and Automation*, 9(6), 297–308. doi:10.14257/ijca.2016.9.6.28
- Wiseman, Y., & Grinberg, I. (2018). The Trolley Problem Version of Autonomous Vehicles. *The Open Transportation Journal*, 12(1). doi:10.2174/18744478018120100105
- Wiseman, Y., Schwan, K., & Widener, P. (2005). Efficient end to end data exchange using configurable compression. *Operating Systems Review*, 39(3), 4–23. doi:10.1145/1075395.1075396
- Wong, J. C. (2018). Rage against the machine: self-driving cars attacked by angry Californians. *The Guardian*. Available online at: <https://www.theguardian.com/technology/2018/mar/06/california-self-driving-cars-attacked>
- Worldometers info. (2018). Available online at: <https://www.worldometers.info/world-population/>
- Yoeli, R. (2002). Ducted fan utility vehicles and other flying cars. *2002 Biennial International Powered Lift Conference and Exhibit*, 5995. 10.2514/6.2002-5995

Zohdy, I. H., Kamalanathsharma, R. K., & Rakha, H. (2012, September). Intersection management for autonomous vehicles using iCACC. *Proceedings of 15th International IEEE Conference on Intelligent Transportation Systems (ITSC 2012)*, 1109-1114. 10.1109/ITSC.2012.6338827

ADDITIONAL READING

Blau, M., Akar, G., & Nasar, J. (2018). Driverless vehicles' potential influence on bicyclist facility preferences. *International Journal of Sustainable Transportation*, 12(9), 665–674. Advance online publication. doi:10.1080/15568318.2018.1425781

Buehler, M., Iagnemma, K., & Singh, S. (Eds.). (2009). *The DARPA urban challenge: autonomous vehicles in city traffic*. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-03991-1

Lozano-Perez, T. (2012). *Autonomous robot vehicles*. Springer Science & Business Media.

Milakis, D., Van Arem, B., & Van Wee, B. (2017). Policy and society related implications of automated driving: A review of literature and directions for future research. *Journal of Intelligent Transport Systems*, 21(4), 324–348. doi:10.1080/15472450.2017.1291351

Rothenbücher, D., Li, J., Sirkin, D., Mok, B., & Ju, W. (2016, August). Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016)*, pp. 795-802. 10.1109/ROMAN.2016.7745210

Wiseman, Y. (2017). Automatic Alert System for Worn Out Pipes in Autonomous Vehicles. *International Journal of Advanced Science and Technology*, 107, 73–84. doi:10.14257/ijast.2017.107.06

Wiseman, Y. (2018). In an Era of Autonomous Vehicles, Rails are Obsolete. *International Journal of Control and Automation*, 11(2), 151–160. doi:10.14257/ijca.2018.11.2.13

Wiseman, Y. (2019). Driverless Cars will Make Passenger Rails Obsolete, IEEE. *Technology in Society*, 38(2), 22–27.

Zhao, L., Ichise, R., Yoshikawa, T., Naito, T., Kakinami, T., & Sasaki, Y. (2015, June). Ontology-based decision making on uncontrolled intersections and narrow roads. In *IEEE Intelligent Vehicles Symposium (IV 2015)*, pp. 83-88. 10.1109/IVS.2015.7225667

Zhou, M., Qu, X., & Jin, S. (2017). On the impact of cooperative autonomous vehicles in improving freeway merging: A modified intelligent driver model-based approach. *IEEE Transactions on Intelligent Transportation Systems*, 18(6), 1422–1428.

KEY TERMS AND DEFINITIONS

Absolute Positioning: A system that is able to locate the precise autonomous vehicle position, so the vehicle will be able to know where it should make its turns.

Automated Highway System (AHS): A system that is capable to drive a vehicle in a highway by itself in a pre-determined route. This system can significantly reduce distances currently necessitated

between vehicles, so as a result such systems will be able to relieve many traffic congestions. Such systems are typically combined with adaptive cruise control and collision avoidance systems and typically used by autonomous vehicles level 3 or higher.

Computer Vision: Systems that are in charge of an autonomous vehicle's ability to observe its surroundings.

Far Infrared Sensors (FIRS): Sensors employed by autonomous vehicles with the purpose of obtaining the heat outline of pedestrians or animals, so as to avoid a collision with them. These sensors can work together with traditional cameras and LIDAR, so as to decide together on the safest choice.

Intelligent Driver Assistance Systems (IDAS): Devices aiming at improving vehicle safety, for example, blind spot monitor or traffic signs recognition.

Internet of Vehicles (IoV): Internet of vehicles is a wireless network used for information exchange between vehicles, infrastructure, pedestrians using the traditional Internet protocols and networks.

Light Detection and Ranging (LIDAR): A laser-based radar used in autonomous vehicles. In nowadays autonomous vehicles, the LIDAR unit is typically located on the vehicle roof with the aim of enabling unimpeded 360-degree view of the autonomous vehicle surroundings.

Platooning: Multiple autonomous vehicles that communicate and exchange information in order to move and maneuver simultaneously. Commonly the vehicles go in a convoy, so as to increase road capacity.

Vehicle to Infrastructure (V2I): Communication technology enabling information exchange between vehicles and infrastructures.

Vehicle-to-Vehicle (V2V) Communication: Communication technology enabling information exchange between vehicles.

Convolutional Neural Network

Mário Pereira Véstias

 <https://orcid.org/0000-0001-8556-4507>

INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal

INTRODUCTION

In a broad sense, a convolutional neural network is one of many methods to achieve artificial intelligence.

Artificial intelligence (AI) is a field of computer science dedicated to the research of methods and algorithms that permit to perceive information from the environment, learn from it and taking actions and decisions based on the learning outcomes without any explicit orientation from external agents. AI consists of many sub-fields, including evolutionary computation (e.g., genetic algorithms), vision (e.g. object recognition), expert systems (e.g. decision support systems), speech processing (e.g. speech recognition), natural language processing (e.g., machine translation), and machine learning (e.g. decision trees).

Machine learning (ML) is a type of algorithms and models for computing systems to do tasks based on pattern identification and inference (Mitchell, T., 2017). When it is difficult or infeasible to develop an algorithm to do a particular task, machine learning algorithms can provide an output based on previous data. Machine learning algorithms identify features from data and build models from these features so that new decisions can be produced based on these models and rules. Examples of machine learning models include deep learning, support vector machines (SVM), genetic algorithms, clustering, dimensionality reduction, artificial neural network, decision tree and others.

In this chapter we are particularly interested in deep learning (DL) (Aggarwal, C., 2018). Deep learning teach computers to learn by example. Deep learning first appeared in 1980, but only recently has been applied in practice because it requires large amounts of labeled data and high-performance computing. Its application is vast including areas like computer vision, speech recognition, language processing, among others.

The most recent deep learning models are based on artificial neural networks (ANN) with a large number of layers. Deep learning networks are first trained using a large set of labeled data in order to learn features from the training data without being explicitly programmed to learn them.

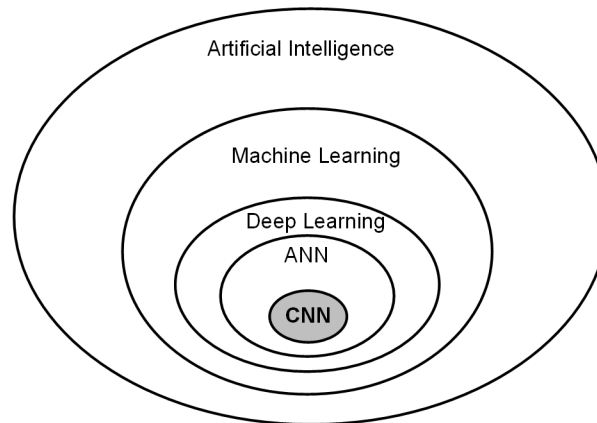
There are several types of artificial neural networks including the feedforward neural network, the Kohonen self organizing neural network, the recurrent neural network, the convolutional neural network, the modular neural network, among others. The convolutional neural network (CNN) is one of the most used deep learning architecture for artificial intelligence (see Figure 1).

Convolutional neural networks are one of the most used deep learning models in the industry and business. Therefore, for those interested in understanding how deep learning is applied in this markets it is fundamental to understand the architecture of CNNs and understand the evolution and trends of CNNs during the last decade.

This article focus on convolutional neural networks with a description of the model, the training and inference processes and its applicability. It will also give an overview of the most used CNN models and what to expect from the next generation of CNN models.

DOI: 10.4018/978-1-7998-3479-3.ch002

Figure 1. Convolutional Neural Networks in the context of artificial intelligence



BACKGROUND

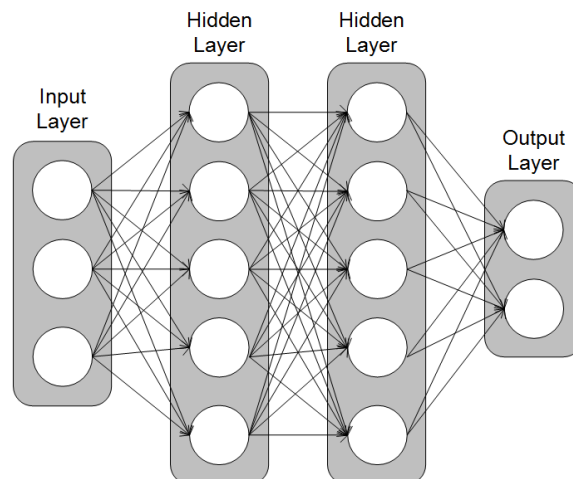
Artificial neural network (ANN) is a machine learning model that mimics the structure of the human brain consisting of interconnected neurons. Theoretically, an ANN is a universal model capable to learn any function (Hornik et al., 1989). A class of ANN is the convolutional neural network introduced in (LeCun, Y. et al., 1989).

Convolutional neural networks process images. The network applies successive convolutions to the input image using several trained filters to identify features of the image. Features identified in a convolution are processed by the next convolutions to identify more complex features. The process permits to learn complex features and at the same time is invariant to translation of the features.

Artificial Neural Network

An artificial neural network consists of input, output and hidden layers, which are all layers between the input and the output layers (see Figure 2).

Figure 2. Artificial Neural Network



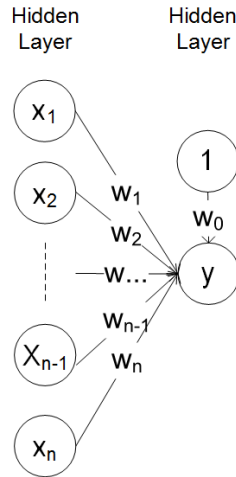
Each layer contains nodes or neurons that map all its inputs into an output to be used by the nodes of the next layer. Nodes of a layer are connected to the nodes of the next layer, except the output layer whose outputs are the result of the neural network. The first layer – input layer – generates the inputs of the first hidden layer and the output layer receives the inputs from the last hidden layer and produces the classification result. A neural network with more than three hidden layers is designated deep neural network (Bengio, Y., 2009).

A layer can have any number of neurons. The input layer has many neurons as the number of input data, while the number of neurons in the output layer depends on the target application. For example, in a neural network used to classify images, the number of output neurons is the number of classification classes of the network.

Neuron Function

Let's consider a neuron with n inputs from n neurons of the previous layer, $\{x_1, x_2, x_3, \dots, x_n\}$. Each connection between neurons has a weight, $\{w_1, w_2, w_3, \dots, w_n\}$ (see Figure 3).

Figure 3. Neuron function



The output of a neuron, y , is computed as follows:

$$y = f\left(w_0 + \sum_{k=1}^n w_k x_k\right)$$

Weight w_0 is in fact an external input called bias introduced in the graph with a unitary neuron. Function $f(\cdot)$ is called an activation function and is used to decide the output of the neuron. Activation functions can be linear or non-linear and are chosen according the domain of application (Nwankpa, C. et al, 2018).

A simple activation function is the Heaviside function that is 0 for negative values and 1 for the others. However, piecewise-linear functions are not differentiable and training requires differentiable func-

tions. A well-known used activation function is the sigmoid given by $\frac{1}{1+e^{-z}}$. It is usually used to predict a probability since it varies between 0 and 1. For multiclass classification, the softmax function is used. This function takes as input a vector of k values and normalizes it into a probability distribution of k probabilities.

Another used activation function is the hyperbolic tangent that varies between -1 and 1, increasing the output range.

The most used activation function is now ReLU (Rectified Linear Unit). The function is 0 when the input is negative and is equal to the input when the input is positive. The function is not differentiable at 0, so it is assumed that the derivate at zero is zero because it gives better results than assuming one.

Training Neural Networks

Training is the process of determining all weights of the network so the outputs of the network give the best mapping results. The training can be supervised or unsupervised. In supervised training, the network knows the output for each input. Unsupervised training is when the network has a training set without the desired outputs. In this case, the network has to find some features of the input data to classify them.

In the supervised training all weights, $Wset$, are adjusted to minimize an objective function that quantifies the error, E , between the M outputs, y_k , and the desired outputs, t_k , for all N data inputs, x_n . Usually, the error is calculated as the sum of squared error:

$$E(Wset) = \sum_{n=1}^N \sum_{m=1}^M (y_k(x_n, Wset) - t_k)^2$$

Training a neural network is an iterative process of two main steps: forward propagation and back-propagation. The first step is to enter an input data and propagate it through the whole network to find the prediction value with a set of weights. Next, a loss function (sum of squared error) is applied to determine the error between the observed output and the expected output. In the ideal case, the loss should be zero, meaning that there is no difference between the observed and the expected values. The loss value is then propagated backwards to all neurons that are in a path from the inputs to the output. Each neuron receives a fraction of the loss according to its contribution to the output. So, neurons with a higher contribution to a particular output will receive a higher fraction of the loss function associated with that output. After all neurons receive the loss fraction, weights are adjusted to reduce the loss. Weights are changed using a technique designated *gradient descent*. According to this technique, weights are changed in small increments based on the derivative of the loss function, that is:

$$\Delta W[i] = -\gamma \left(\frac{\partial E_n}{\partial w[i]} \right)$$

where γ is the learning rate.

Weight initialization is important in the final quality of the trained network. Some works indicate that better results are obtained when the weights are randomly chosen between specific normalized ranges (Glorot, X. & Bengio, Y., 2010).

In the unsupervised training, given a training set without expected outputs, the method has to train the network based on characteristics of the input data. One known method for unsupervised training is the auto-encoders. This technique is similar to a compression algorithm that uses a smaller subset of features to represent the input data (Bengio, Y., 2009).

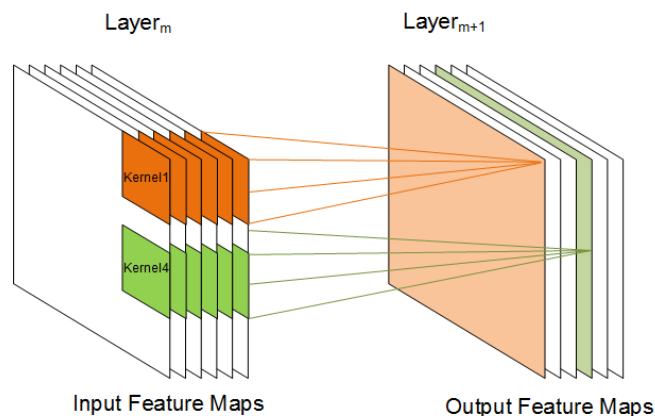
Convolutional Neural Network

A traditional neural network can be applied to images. However, even with images of medium size, the number of input neurons is high and consequently there would be a high number of weights from the input neurons to the first hidden layer. For example, considering an image of size 224×224 , there are over 50000 input nodes. With the same nodes in the second layer, we would need 50000×50000 weights just for the first layer. This number of weights is very difficult to train with good results and needs a huge amount a memory and computational effort. A better approach is to take into consideration the type of input data, an image in this case (LeCun, Y., 1989). Convolutional neural networks consider the spatial information between pixels, that is, the output of a neuron of the input layer is the result of the convolution between a small subset of the image and a kernel of weights.

A regular convolutional neural network has different types of layers: convolutional, pooling and fully connected. CNN may have other particular layers different from these and are known as irregular CNN. In the following, we explain in more detailed the three types of regular layers.

The **convolutional layer** receives a set of input feature maps (IFM) and generates a set of output feature maps (OFM). A feature map is a 2D matrix of neurons and several feature maps form a 3D volume of feature maps (see Figure 4).

Figure 4. Input and output feature maps of a convolutional layer



A 3D block of weights – kernel or filter – is convolved across the width and the weight of the input feature maps and computes the dot product between the kernel and the feature maps generating an output feature map. Each different kernel produces one output feature map. Therefore, the number of output feature maps is the same as the number of filters of the layer. Each kernel is trained to detect a particular feature of the input data. To explore spatial local correlation between adjacent neurons the convolution window (width $\times$ weight) of the kernel is usually small (e.g., 3×3 , 5×5 , etc.), while the depth of the kernel is the same as the number of input feature maps.

Usually, kernels slide across the feature maps shifting one neuron each time (stride of 1). This generates high overlapping that increases with the window size and output feature maps with the same size. Instead, a larger stride can be used resulting in less overlapping and a smaller output volume. Another parameter of the convolution process is the padding of the IFMs. If we run a kernel with window size 3×3 over an input feature map of size $k \times k$, the output feature map will have a size $(k-2) \times (k-2)$. In many cases, it is desirable to preserve the size of the maps. One solution consists to pad the border of the IFMs with zeros. In this case, padding the border with one level of zeros would produce output feature maps with the same size of the input feature maps.

The **pooling layer** subsamples the IFMs to achieve translation invariance and reduce over-fitting. Basically, the relative location of a feature is more important than its absolute location.

Given an input image and the pooling size (typically two or three), pooling partitions the image in 2D arrays of the size of the pooling window. Then, it applies a pooling function to all members of the pooling window (see Figure 5).

Figure 5. Input and output feature maps of a convolutional layer



Examples of pooling functions are average pooling (average between the output neurons of the window) and max pooling (maximum value between the output neurons of the window). Considering, for example, a pooling window with entries $\{2, 4, 6, 8\}$, the max pool would return 8, while average pooling returns 5. Max pooling gets faster convergence during training (Scherer, D. et al, 2010). Also, it reduces the number of weights and consequently the required memory to store them and the amount of computation. Pooling is not applied to all convolutional layers.

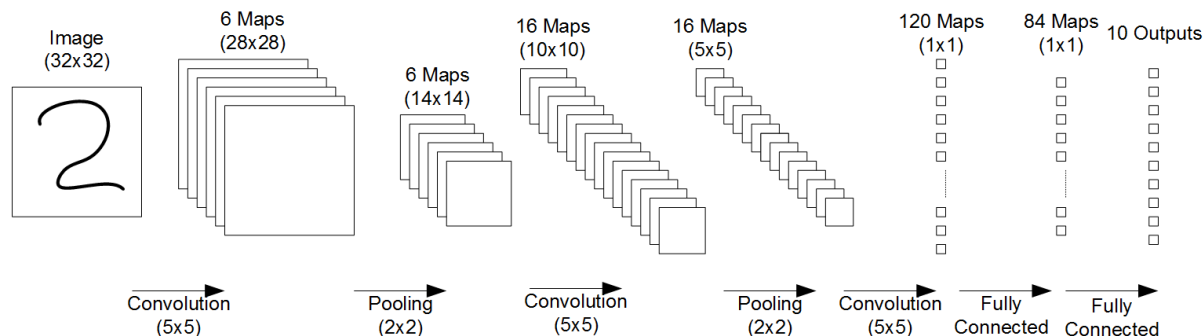
The size of the pooling window determines the reduction of the feature maps. The typical size of 2×2 applied with a stride of two downsamples the input maps in both dimensions reducing the original map 75%. Note that pooling reduces the size of the map but the number of feature maps remains the same.

The last layers of a convolutional network are **fully connected**. A fully connected layer follows the structure of the traditional neural network where all neurons of the layer are connected to all neurons of the previous layer. Algorithmically, it is a vector-matrix multiplication followed by vector addition of the bias values.

To better understand the concepts introduced, let's consider the first convolutional neural network to achieve high accuracy for hand digit classification LeNet (LeCun, Y. et al., 1995). LeNet has two sets of convolutional and average pooling layers, followed by a flattening convolutional layer, two fully connected layers and a softmax classifier. The activation function is the hyperbolic tangent (see Figure 6).

The input of LeNet is a grayscale image of size 32×32 . The first convolutional layer uses six filters with a window size of 5×5 and a stride of one producing six output feature maps of size 28×28 , since it does not consider padding. Then it is applied an average pooling layer with a filter of size 2×2 and a stride of two, reducing the feature maps to 14×14 . A second convolutional layer is applied with sixteen

Figure 6. Architecture of LeNet



filters of size 5×5 and stride of one generating sixteen output feature maps of size 10×10 which are then pooled to size 5×5 . The next layer is a convolutional layer with 120 feature maps of size 1×1 . Then comes one fully connected layer with 84 units followed another fully connected layer with ten outputs, one for each digit and the softmax classifier.

Recent Convolutional Neural Networks

CNNs have many parameters to be defined, like the number of layers, the number of neurons, the size of filters, the size of pooling, the activation function, the learning rate, more and new type of layers, among others. The exploration of design space defined by all these parameters lead to many proposals of CNN. In the following, some of the most known CNN for its results and novelty will be described.

LeNet

LeNet (LeCun, Y. *et al.*, 1995) was able to demonstrate some robustness features of CNN to rotation, distortion, scale, etc. In total, the network has 60K parameters or weights and an accuracy above 99% (see architecture in Figure 5, in the previous subsection).

AlexNet

The next big step of CNN application was achieved with AlexNet (Krizhevsky, A., 2012) considered the first CNN with great results for image classification. AlexNet is a deeper network than LeNet and 60 M parameters, 1000× more parameters than LeNet. AlexNet has five convolutional layers, three pooling layers after convolutional layers two, three and five, and three fully connected layers. Instead of the hyperbolic tangent, the activation function of AlexNet was ReLU to improve the convergence rate of learning and the problem of vanishing gradient (Hochreiter, S., 1998). The input images have size $227 \times 227 \times 3$ and the network uses larger filter sizes in in the first (11×11) and second layers (5×5).

AlexNet won the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) with a top-5 error rate of 17.0% (top-5 is the error rate at which given an image the network does not include the correct label within its top five predictions) and a top-1 error rate of 37.5% with ImageNet (Imagenet is a large database of images for use in object recognition research. It has over 15 million annotated images with more than 22000 categories).

ZefNet

Until 2013, the design of a CNN was based on trial and error, changing the parameters without a clear understand of what was happening in the CNN layers. In 2013 a multiplayer deconvolutional neural network was proposed in (Zeiler, M. D. & Fergus, R., 2013) named ZefNet. Instead of just proposing a new CNN, the authors also proposed a technique to observe the network activity of AlexNet (can be applied to other networks) named Deconvolutional Network, which monitors the output of neurons to establish correlations between features and input neurons. The idea has already been explored with Deep Belief Networks (Erhan, D., 2009) and Autoencoders (Le, Q. V., 2013).

When applied to AlexNet the method have shown that only a few neurons were active in the first two convolutional layers. Based on these observations, the authors have fine tuned AlexNet achieving a top-5 error rate of 11.2%. A few more modifications in both the model and the training process were considered. For the training process, they considered only 1.3 million images, but even so it took twelve days to train the network. They substituted the filter of the first convolutional layer (size 11×11) by a filter of size 7×7 to retain more information from the image.

VGG

In 2014 the VGG neural network brought some new directions to CNN development. They considered two main foundations for the development of VGG, simplicity and uniformity of layers and depth of the network. Six different architectures were tested and the one with best accuracy was VGG-16 with sixteen layers using only 3×3 filters with stride and pad of 1, and 2×2 pooling with stride 2. The authors concluded that two convolutional layers with 3×3 filters has the same results of a single layer with 5×5 filters. The same for three 3×3 filters instead of a single 7×7 filter. This reduces the number of parameters and applies the activation function more times. Another option in the design of VGG was to increase the number of filters after each pooling step as we move to the next layers. So, the size of the feature maps decrease but the depth increases as we move along the network.

Like the previous architectures it was trained with batch gradient descent and the training took from two to three weeks with four GPUs. VGG did not win the ImageNet contest that year but improved previous year top-5 error rate to 7.3%. The main drawback of VGG is that it has 138 million parameters, the main reason for its long training times.

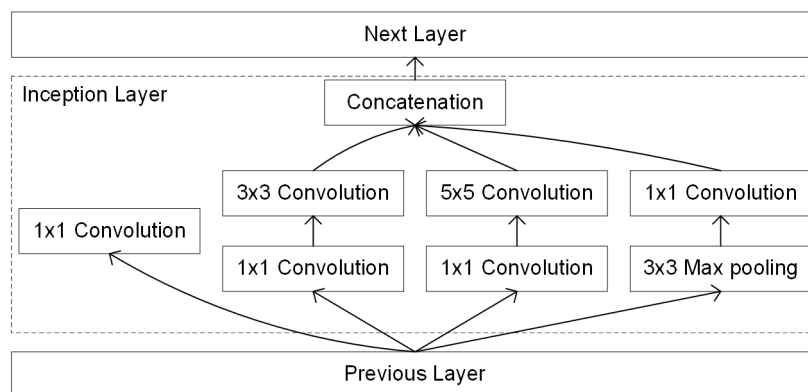
GoogleNet

The winner of the ILSVRC in 2014 was GoogleNet (Szegedy, C. *et al.*, 2014), also known as Inception-V1. Previous CNNs followed a regular approach of stacking convolutional, pooling and fully connected layers with variations in the number of layers, and number and size of filters. GoogleNet broke this regular structure. The layers of a regular CNN are serialized with one after the other. In GoogleNet groups of convolutions run in parallel defining a module named Inception (see Figure 7).

In the inception module, there are four main operations in parallel, namely three convolutions with different sizes (1×1 , 3×3 , 5×5) and one 3×3 max pooling. To reduce the number of operations, the 3×3 and 5×5 convolutions are preceded by 1×1 convolutions to reduce the volume of data.

The inception module is like a small network inside a larger network that is able to extract information at different scales with different window sizes. The pooling operation reduces the size of output data.

Figure 7. Inception module of GoogleNet



The complete GoogleNet network has nine inception modules and does not use fully connected layers. Instead, an average pool is used after the last inception that reduces the output $7 \times 7 \times 1024$ volume of the last inception module to a $1 \times 1 \times 1024$ volume.

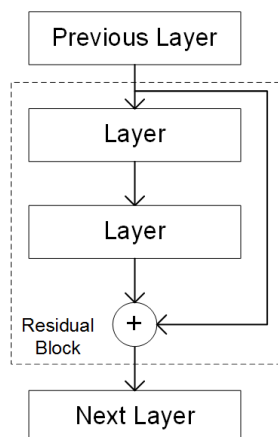
GoogleNet uses 6.8 million parameters to achieve a top-5 error rate of 6.7% for ImageNet.

Other versions of the inception module have followed. Inception-v2 and inception-v3 were presented in the same paper (Szegedy, C., 2016). In this work, the convolution with size 5×5 was replaced by two 3×3 convolutions. They factorize convolutions of size $n \times n$ to a combination of two filters $1 \times n$ and $n \times 1$ applied in parallel. Inception-v4 introduced some more modifications and optimizations. Modules were made more uniform and the complexity was reduced. In spite of uniformization, they created three different inception modules.

ResNet

ResNet (He, K., et al., 2015) has increased the number of layers to 152 and won the ILSVRC competition in 2015 with a top-5 error rate of 3.6% (human error for image classification is from 5 to 10%). ResNet not only pushed the depth of the network but have also introduced a new block named *Residual Block* (see Figure 8).

Figure 8. Residual Block of ResNet



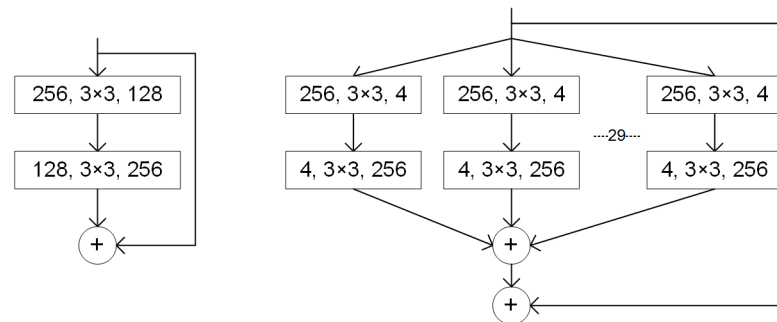
In the residual block the result a series of two convolutional layers is added to the input of the residual block. So, instead of just processing the original input with a series of convolutional layers, the result of the convolutions is added to the original input, like a delta factor. According to the paper, the new structure permits a better optimization of the network.

Similar to what happened in VGG, ResNet increases the layers and decreases the size of maps very soon. After the first two layers maps are reduce to one quarter of the original image size.

ResNeXt

ResNeXt (Xie, S., et al., 2017) is the follower network of ResNet with slightly better results than the original ResNet with a top-5 error of 3.03%. The model introduces the concept of cardinality that splits the original residual block without increasing the complexity (see Figure 9).

Figure 9. Residual block with cardinality



On the left, we have an example of a residual block with two convolutions and on the right we have a cardinality block that splits the original residual block in 32 blocks (the cardinality of the block). The complexity is the same since the set of filters is also split across each member of the cardinality block.

DenseNet

DenseNet (Huang, G., 2018) is a CNN architecture also based on ResNet with some modifications in the residual block. The work proposes the *Dense Block* to replace the residual block. The dense block has several layers, but instead of receiving data only from the previous layer it receives data from all preceding layers and sends the output to all subsequent layers within the block. Instead of addition, the block uses concatenation to group data from previous blocks.

The structure has some advantages over ResNet. During training the backward error propagation reaches earlier layers with a straight path. The number of parameters in DenseNet is smaller than in ResNet (about half for the same number of layers). Receiving information from several previous layers increases the diversity of features processed by a single layer. The accuracy of DenseNet is close to the one achieved with ResNet but with less parameters.

SENet

SENet (Hu, J., Shen, L., Sun, G., 2018) introduced a new network block – squeeze-and-excitation. The idea of the block is to emphasize important features and cancel less useful ones. The block runs the convolutions of the residual block of ResNet. Each feature map is then squeezed to a single value using a global average pooling. Next, a fully connected layer followed by ReLU is applied to introduce non-linearity. A second fully connected layer is applied followed by a sigmoid activation function. The result is finally added to the original input, like in the residual block.

With these modifications, SENet won the ILSVRC competition in 2017 with a top-5 error rate of 2.25%.

Application Examples of Convolutional Neural Networks

The set of target applications of convolutional neural networks is growing. Since these networks are dedicated to image analysis, any application associated with image and video is a potential candidate for this type of deep neural networks. Face recognition (Kalinowski, I., & Spitsyn, V., 2015) and scene labelling (Pinheiro, P. & Collobert, R., 2014) already applies successfully a CNN. Face recognition with a CNN identifies faces in an image and match each face to a name from a database. The advantage of the CNN model compared to other machine algorithms is its high accuracy independent of external factors, such as light or view angle. Face recognition is applied in social networks, like Facebook, entertainment applications that change the aspect of the face, surveillance cameras, among others.

Other type of image recognition is the process of identifying written characters of symbols from an image. Image tagging is also using CNNs. eCommerce platforms are using these applications to match the images accessed by users to their commercial needs.

Medical image analysis is another hot area that is starting using CNNs to analyze medical images (Razzak M.I., 2018). Images associated with many types of exams are classified with a CNN with an accuracy higher than that obtained by the human. With the vast amount of medical information available, there is enough data for CNNs to analyze medical images with high success. Medicine applications are not just image analysis. Health risk assessment is another application area, where CNNs are used to analyze the past data of a patient and determine the probability of occurrence of certain diseases, propose the best way to treat it or even explore the risks for people close to the patient.

Autonomous driving is another hot topic for convolutional neural networks (Kocic, J., 2019). The system receives images from cameras attached to the vehicle and determines the driving direction of the vehicle. The task is complex since the vehicle has to take into consideration many other aspects, like traffic sign recognition, detection of persons in the street, etc.

Convolutional neural networks are applied in many other machine learning applications. The list is already vast, including code generation (Pengcheng, Y. & Neubig, G., 2017), analysis of sentiments (Lei, T., Barziley, R. & Jaakkola, T., 2016), image segmentation (He, K., et al., 2017), text recognition (Jaderberg, M., et al., 2016) and object recognition (Redmon, J., & Farhadi, A., 2017), just to name a few.

FUTURE RESEARCH DIRECTIONS

Many CNN models are proposed every year and the accuracies are increasing with each network. Initial CNN were regular structures based on convolutional, pooling and fully connected layers. The accuracy

of CNN was then improved with the introduction of new layers. These irregular structures have improved the training and inference of CNN at the cost of some specificity of particular layers.

In spite of all the evolution, the study and development of CNNs are still in its infancy and new challenges are yet to be resolved. Some of these challenges are associated with training. CNN are very good on classifying good images but are still little robust to image distortions and noise. Another problem is that it is difficult to understand and verify what is going inside the layers and the influence of each filter or layer over the network results. Another major challenge is that CNNs require large amounts of data to learn. So, those applications with few training images are difficult to be modeled with a CNN. Training a CNN is a hard processing task that requires high-performance computing platforms. Inference is less demanding but still requires high computing power. How to do inference and training in edge devices with low computing resources is an actual area of intense research.

CNN still continue to evolve and new achievements are expected in the near future. A recent training issue is how to incrementally train an already trained network. A running CNN can keep learning with new data to improve its classification capability. The question is how to incrementally change the network without retraining the network for all input data set.

The time needed to train a network is a major concern since it limits the continuous increase of network complexity. Also, the inference in edge devices with low resources also limits the network complexity. Two paths have been followed to overcome this embedded limitations. One way is to improve hardware computing power and apply network optimizations to reduce complexity. Another way is to design networks for embedded devices (Sandler, M.B., 2018). So, CNN development cannot be independent of the target hardware.

Many parameters have to be determined in a large CNN. Many of these, like size and number of filters, are explored manually. Some automatic algorithms can be used to explore the design space but it takes a long time. So, it is important to clearly understand the relation between each type of parameters and the network accuracy so that the design process can be faster.

CONCLUSION

The article describes the fundamentals of convolutional neural networks, the most known types of layers and CNNs proposed in the literature. Since the first successful CNN for character recognition, followed by AlexNet for image classification that many new networks have been proposed with better accuracies. Networks with more layers, new types of layers, new optimization techniques to help the accuracy and improve training have helped to evolution of CNN.

Researchers have been concerned about understanding the internals and mechanisms of CNN so that new improvements have some reasoning and are not just based on an empirical process of trial-and-error. Another concern is about the complexity of the networks and the computing requirements. In a near future it is expected an increase of CNNs running in edge devices. This requires an extra care over the development of CNN since they have to run in a constrained computing platform.

Therefore, in spite of 20 years of research and development, CNNs are still an actual topic of intense research.

REFERENCES

- Aggarwal, C. (2018). *Neural Networks and Deep Learning: A Textbook* (1st ed.). Springer. doi:10.1007/978-3-319-94463-0
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 1–127. doi:10.1561/22000000006
- Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *Univ. Montr.*, 1341, 1.
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Artificial Intelligence and Statistics, International Conference on*, 249–256.
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *International Conference on Computer Vision (ICCV)*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Multimedia Tools and Applications*, 77, 10437–10453.
- Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 6(02), 107–116. doi:10.1142/S0218488598000094
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. doi:10.1016/0893-6080(89)90020-8
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141.
- Huang, G., Liu, Z., Maaten, L., & Weinberger, K. (2018). Densely Connected Convolutional Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2016, January). Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*, 116(1), 1–20. doi:10.1007/11263-015-0823-z
- Kalinowski, I., & Spitsyn, V. (2015). *Compact Convolutional Neural Network Cascade for Face Detection*. CoRR, abs/1508.01292
- Kocic, J., Jovicic, N., & Drndarevic, V. (2019). An End-to-End Deep Neural Network for Autonomous Driving Designed for Embedded Automotive Platforms. *Journal of Sensors*, 19, 9. PMID:31058820
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1–9.
- Lawrence, S., Giles, C., Lee, T., Ah, C., & Back, A. (1997). Face Recognition: A Convolutional Neural Network Approach. *Neural Networks, IEEE Transactions on*, 8(1), 98–113.
- Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 8595–8598. 10.1109/ICASSP.2013.6639343

LeCun, Y. (1989). Generalization and network design strategies. *Connectionism in Perspective*.

LeCun, Y. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks Stat. Mech. Perspect.*, 261, 276.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551. doi:10.1162/neco.1989.1.4.541

Lei, T., Barzile, R., & Jaakkola, T. (2016). Rationalizing Neural Predictions. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. 10.18653/v1/D16-1011

Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5–6), 555–559. doi:10.1016/S0893-6080(03)00115-1 PMID:12850007

Mitchell, T. (2017). *Machine Learning* (1st ed.). McGraw Hill Education.

Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S. (2018). *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*. Corr, abs/1811.03378

Pengcheng, Y., & Neubig, G. (2017). A Syntactic Neural Model for General-Purpose Code Generation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1, 440–450.

Pinheiro, P., & Collobert, R. (2014). Recurrent convolutional neural networks for scene labeling. *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 32.

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In N. Dey, A. Ashour, & S. Borra (Eds.), *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics* (Vol. 26). Springer. doi:10.1007/978-3-319-65981-7_12

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525. 10.1109/CVPR.2017.690

Sandler, M. B., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. 10.1109/CVPR.2018.00474

Scherer, D., Müller, A., & Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. *Artificial Neural Networks, International Conference on*, 92–101. 10.1007/978-3-642-15825-4_10

Szegedy, C. (2014). *Going Deeper with Convolutions*. arXiv:1409.4842

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. 10.1109/CVPR.2016.308

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10.1109/CVPR.2017.634

Zeiler, M. D. & Fergus, R. (2013). *Visualizing and Understanding Convolutional Networks*. arXiv Prepr. arXiv1311.2901v3

ADDITIONAL READING

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer Verlag.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11, 625–660.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554. doi:10.1162/neco.2006.18.7.1527 PMID:16764513

KEY TERMS AND DEFINITIONS

Activation Function: The activation function defines the output of a neuron given a set of inputs from the previous layer or data input.

Artificial Neural Network (ANN): It is a computing model based on the structure of the human brain with many interconnected processing nodes that model input-output relationships. The model is organized in layers of nodes that interconnect to each other.

Convolutional Layer: A network layer that applies a series of convolutions to a block of input feature maps.

Convolutional Neural Network (CNN): A class of deep neural networks applied to image processing where some of the layers apply convolutions to input data.

Deep Learning (DL): A class of machine learning algorithms for automation of predictive analytics.

Deep Neural Network (DNN): An artificial neural network with multiple hidden layers.

Feature Map: A feature map is a 2D matrix of neurons. A convolutional layer receives a block of input feature maps and generates a block of output feature maps.

Fully Connected Layer: A network layer where all neurons of the layer are connected to all neurons of the previous layer.

Network Layer: A set of neurons that define the network of a CNN. Neurons in a network layer are connected to the previous and to the next layer.

Pooling Layer: A network layer that determines the average pooling or max pooling of a window of neurons. The pooling layer subsamples the input feature maps to achieve translation invariance and reduce over-fitting.

Softmax Function: A function that takes as input a vector of k values and normalizes it into a probability distribution of k probabilities.

Supervised Training: A training process of neural networks where the outcome for each input is known.

Unsupervised Training: A training process of neural networks where the training set does not have the associated outputs.

Human-Agent-Robot Teamwork (HART) Over FiWi-Based Tactile Internet Infrastructures

1

Mahfuzulhoq Chowdhury

Institut National de la Recherche Scientifique (INRS), Canada & University of Quebec, Canada

Martin Maier

Institut National de la Recherche Scientifique (INRS), Canada

INTRODUCTION

In the current digital Darwinism era, technologies are evolving faster than the business organization can naturally adapt (Daugherty, 2018). In the past, business executives focused on using machines to automate specific workflow processes. Traditionally, these processes are linear, stepwise, sequential, and repeatable. But performance gains from that traditional approach have recently been leveling off, as companies wring the last bits of efficiencies from mechanistic automation. The traditional approach (either human-only activity or machine-only activity) has been to view humans and machines as rivals, each side fighting for the other's job. The truth is that companies can achieve significant boosts in performance, when machines and humans work together as allies, not adversaries, to capitalize on each other's complementary strengths. For instance, performing routine tasks and detecting hidden patterns in an image can be easy for machines, whereas dealing with unsatisfied customers can be easy for humans. To achieve higher productivity gains than traditional systems, currently leading firms in many industries are re-imagining their business processes to be more flexible, faster, and adaptable to customer behavior, preferences, and needs of their workers in a given moment. This trend of exploiting the adaptive system for symbiotic collaboration between humans and machines is unlocking what is known as the third wave of business process transformation. This collaboration allows weak workers to transfer tedious tasks to powerful machines and by enabling them to perform their work faster and more efficiently through expert guidance, advice, and support from artificial intelligence (AI) based systems. Hence, to realize beneficial human-machine coactivity in different business processes, the fundamental question that naturally arises is how to enable real-time collaboration and communication between humans and machines. To answer this question, some of the key technological terms are described below in greater detail:

To unleash the full potential of many real-time cyber-physical systems (CPSs) that harness human-machine interaction, including virtual and augmented reality, an extremely low round-trip latency of 1-10 milliseconds is required to match human interaction with the environment. This vision of ultra-low latency communications gives rise to the so-called *Tactile Internet*, which has recently emerged to steer/control virtual and physical objects (e.g., remote-controlled robots) and to transmit touch and actuation in real-time (Maier, 2016). Key Tactile Internet applications include real-time gaming, wireless controlled exoskeletons, remote surgery, elderly/disabled people care, driver-assistance systems to support humans in arriving safely and comfortably at their destinations. By enabling ultra-low latency communications between humans and machines in conjunction with carrier-grade robustness and availability, the Tactile

DOI: 10.4018/978-1-7998-3479-3.ch003

Internet represents a paradigm shift from traditional wired and mobile Internet based content-delivery to labor-delivery networks via tactile/haptic devices. At the nexus of computerization, automation, and robotization lies the emerging Tactile Internet, which will be centered around human-to-machine/robot (H2M/R) communications by providing the medium for transporting haptic (i.e., touch and actuation) and non-haptic (data, video, and audio) traffic in real-time (Maier, 2018). By offering highly reliable and responsive network connectivity for real-time human-machine interaction centric applications, the Tactile Internet holds great promise to have a profound socio-economic impact on our everyday life, ranging from healthcare (e.g., remote robotic surgery), transport (e.g., driver assistance system) to entertainment (e.g., online gaming) and manufacturing industry. The long-term goal of the Tactile Internet is to enable emerging products and services that complement humans rather than substitute for them. Recently, Stanford University launches its inaugural report on “*Artificial Intelligence and Life in 2030*,” in which they predicted that within next 10-15 years a vast amount of intelligent human-aware systems will be developed. Unlike the Internet of Things (IoT) without any human involvement in its underlying machine-to-machine (M2M) communications, the Tactile Internet involves the inherent human-in-the-loop (HITL) nature of haptic interaction and thus allows for a human-machine cooperative design approach towards creating and consuming novel immersive experiences via the Internet.

Recently, Garry Kasparov emphasized the need for a strong coordination process for better human-machine co-activity in his book “*Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*,” where he shows that a strong human worker + machine worker + weak coordination process is less effective than its weak human worker+ machine worker+ strong coordination process counterpart. In other words, an intelligent process with inferior technology and inferior knowledge outperforms a less intelligent process with superior technology and knowledge. This observation of Garry Kasparov was recently also appreciated by Google and other Silicon Valley companies. In fact, leading technological companies have been putting more efforts in developing AI systems that augment human capabilities (see, e.g., IBM’s Watson) instead of replacing them with autonomous systems. According to Kasparov, this isn’t just user experience (UX), but entirely new ways of bringing human-machine coactivity into diverse fields (e.g., business process and organization management) and creating the new tools/intelligent process we need in order to do so.

A promising approach toward achieving advanced human-machine coordination by means of a superior process for fluidly orchestrating human and machine coactivity may be found in the so-called *human-agent-robot teamwork (HART)* research (Johnson, 2011). In HART, the dynamic allocation of functions and tasks between humans and machines (e.g., robots and cloud servers) is crucial, which may vary over time or be unpredictable in different situations. HART differs from the traditional humans-are-better-at/machines-are-better-at (HABA/MABA) approach, which only allows human-only activity/machine-only activity without driving any symbiotic human-machine development in search for synergies (Bradshaw, 2012). Conversely, with a HART-centric Tactile Internet design approach, humans, collaborative robots, remote cloud, and decentralized cloudlet resources together acting as an intelligent multi-agent system exploit the different characteristics of requested tasks and jointly execute them by using smart orchestration techniques. Considering the underlying human-machine interaction in HART-centric Tactile Internet applications, the main challenge is to orchestrate how tasks can be best shared in concert. Collaboration and communication among HART members are essential to cope with dynamic changes in the task environment, thereby improving the task execution latency. Note that depending on location awareness, a HART task can be either a location oriented physical task (e.g., sensing at a given task location, moving and carrying products to another location), a location independent cognitive task (e.g., caching or object recognition, which might be offloaded for computation onto a powerful cloud server/agent

for processing), or may include both location dependent and independent tasks (e.g., sensing at a task location, sensed object recognition, and cached content delivery of the sensed object). Note that due to its movement and computation capabilities, a collaborative robot may perform both physical and cognitive sub-tasks. Conversely, a cloud agent (remote cloud or local cloudlet) can perform only offloaded, location independent cognitive sub-tasks that require computation and processing resources. We note, however, that due to their higher computation processing and storage capabilities, cloud agents are more suitable for cognitive sub-task processing than resource-limited collaborative robots. Thus, the selection of suitable actors for task assignment is key to reduce the incurred computation and communication overhead during the execution of HART-centric tasks. Moreover, the interdependent activities of HART members may lead to an increased complexity and resource consumption. To facilitate cost-effective HART task execution, research in the area of decentralized network coordination, adaptive bandwidth resource management with traditional broadband and offloading traffic co-existence, real-time task assignment, monetary, time, and energy cost-aware task offloading strategies as well as collaborative computing, mobility management, and failure avoidance strategies become mandatory.

To execute real-time HART-centric applications in a cost-effective manner, network operators have been looking for reliable, fast, low-cost, and future-proof communication infrastructures. Towards this end, integrated *fiber-wireless (FiWi) access networks* that combine the high capacity and reliability of optical fiber networks with the ubiquity and mobility of wireless networks represent promising network infrastructures (Maier, 2015). FiWi networks are based on optical fiber (Ethernet passive optical network) and wireless Ethernet (WLAN) technologies, which are then unified with their cellular counterparts, to give rise to the so-called FiWi enhanced Long Term Evolution Advanced (LTE-A) heterogeneous networks (HetNets). FiWi enhanced mobile networks offer several benefits for low-latency HART task execution. First, they exploit the locality of different HART task and cloud agent/robot resources for local/non-local HART task execution. Second, placing a decentralized resource coordinator at the optical-wireless interface of FiWi enhanced mobile networks may play a major role in avoiding higher delays due to task request routing and task allocation to suitable HART members/actors, including their re-allocation in case of failures.

FOCUS OF THE ARTICLE

For beneficial human mobile device and collaborative machine (robot and cloud agent) coactivity-based HART task execution, this chapter first discusses the related work and open challenges in Background section. This section also identifies several key technologies and limitations of existing actors (i.e., robot, cloud agents) selection schemes for MU's device HART task execution. To tackle some of the existing challenges, this work then proposes a latency sensitive HART task migration scheme in the section of proposed HART task coordination that efficiently assigns tasks among human mobile users (MUs), central and decentralized computational agents (cloud/cloudlets), and collaborative robots across converged FiWi network infrastructures. Specifically, to satisfy the task execution requirements, this section of the book chapter investigates when and where (e.g., robot and cloud agent) a MU's computation-intensive task is migrated for execution considering task characteristics, actors (robot and cloud server), bandwidth resources availability, and users mobility. Further, this section describes a bandwidth allocation scheme that allocates timeslot resources to MU's for their broadband and task migration traffic transmission. Furthermore, this section presents an analytical model that analyzes both predicted task execution time and power consumption value of proposed scheme. In evaluation results and findings section, this work

examines the numerical performances of our proposed migration scheme in comparison with traditional baseline schemes. Future research direction section outlines open research issues for HART task coordination over FiWi enhanced Tactile Internet infrastructures. Finally, in the conclusions section, this book chapter summarizes our findings.

BACKGROUND: RELATED WORKS AND RESEARCH CHALLENGES

H2R/M communication based Tactile Internet application execution sets demanding requirements for future access networks in terms of latency, reliability, and capacity. To achieve the Tactile Internet's key requirements of very low latency and ultra-high reliability, recently in (Beyranvand, 2017), the authors proposed the concept of so-called FiWi enhanced LTE-A HetNets that unify coverage-centric 4G mobile networks and capacity-centric FiWi broadband access networks based on data-centric Ethernet technologies. The authors showed that an average end-to-end latency of 1 millisecond and almost guaranteed availability can be achieved in FiWi enhanced LTE-A HetNets by means of fiber backhaul sharing and WiFi offloading capabilities. Note, however, that only conventional human-to-human (H2H) communications was considered in (Beyranvand, 2017), without any coexistent H2R/M or M2M communications. To realize low-latency H2R communications in the Tactile Internet, in (Maier, 2016) the authors discussed the role of several key enabling technologies, including FiWi enhanced LTE-A HetNets, cloudlets, cloud robotics, and mobile-edge computing (MEC), among others. To extend the capabilities of both tele-operated and networked robotics for different Tactile Internet applications, they also elaborated on the importance of a cloud robotic architecture that leverages the combination of an ad-hoc cloud formed by M2M communications among participating robots and an infrastructure cloud enabled by machine-to-cloud (M2C) communications between robots and the remote cloud. Moreover, by leveraging on the improved throughput, reliability, and delay performance of FiWi enhanced LTE-A HetNets, the authors reported that MEC/cloudlet enhanced FiWi multi-robot infrastructures will be essential for the coordination of low-latency Tactile Internet applications.

To offer enhanced quality of service for the execution of HART-centric Tactile Internet applications, in (Chowdhury, 2017a) the authors discussed the potential of suitable human-to-robot task allocation strategies, which emerged as an interesting research problem, taking a wide variety of task (e.g., task location, workload), robot capabilities (CPU speed, moving speed, remaining energy), and failure during task execution into account. The authors indicated that suitable robot selection schemes are mandatory to reduce the task execution latency and energy consumption of MUs and robots. At present, research in the area of H2R task allocation is in its infancy. In (Trigui, 2014), the authors proposed a robotic task allocation scheme that considers only a robot's distance to a given task location for suitable robot selection. A decentralized task allocation framework was proposed in (Mendonça, 2013), where task requests with higher priority are allocated before lower-priority tasks, though task re-allocation and failure avoidance were not studied. In (Choudhury, 2009), the authors presented a role-based robot selection mechanism that depends on the matching of the task type and the robot's skill. Most of the aforementioned studies on suitable robot selection focused mainly on the involved robots' task processing power or distance. To render the HART-centric task allocation more effective, additional robot properties (e.g., availability, skill, mobility, minimum distance to task location, and energy consumption) need to be considered.

Despite ongoing developments in the robotic industry, the selected robot may not always satisfy given task deadline requirements of many computation-intensive tasks due to limited resources (e.g., CPU speed, storage, or battery power). To overcome these problems, robots may utilize the resources

of other collaborative HART members, e.g., cloud servers/agents. Mobile cloud computing (MCC) has emerged as a potential solution for resource limited robots/mobile devices, which allows mobile devices to offload their computation-intensive tasks fully or in part to resource rich surrogates/cloud servers for processing (Deng, 2015). Note that task offloading to remote cloud servers may not always improve a mobile device's task execution latency due to increased wide area network (WAN) delays. To overcome this burden, the concept of MEC has emerged, which allows mobile devices to offload computation intensive tasks to nearby edge cloud servers/cloudlets for processing. To reduce the communication latency of remote cloud servers, cloudlets may be placed at the network edge (e.g., base stations) in close proximity to MUs (Maier, 2016). Different from data offloading, the computation task offloading life cycle involves the following activities: task input data uploading from mobile device to surrogate/cloud server, offloaded task processing (computation workload) by surrogate/cloud server, and task result downloaded by the mobile device from the surrogate. Hence, for better task execution performance, the decision of whether or not to offload the task to a cloud server depends on several important factors such as why offloading is required (e.g., improve task execution time or save energy), who wants to offload the task (e.g., robot, mobile device, laptop computer), offloading environment (e.g., static, dynamic), different task properties (e.g., workload), offloading infrastructure (remote/edge cloud computing), and bandwidth, among others.

Toward this end, taking the idea of task offloading a step further, task migration has emerged as a promising option to improve the quality of experience (QoE) of MUs, whose location changes due to mobility (Gkatzikis, 2013). Task migration broadens the scope of conventional computation task offloading by not only transferring the task from an MU to the cloud server, but also from one cloud server to another one for execution. Task migration is beneficial for low-latency applications only if the anticipated task execution time at the new cloud server is smaller than that of the previous server. Hence, the question of how and where an MU's task should migrate is an interesting research challenge for low-latency HART-centric task execution. To answer this question, several aspects need to be considered, e.g., information about the state of the current host and tentative destination server, the number of tasks running on each server, and the incurred task migration latency.

At present, only a few studies exist on task migration that exploit cloud agents for MUs' migrated task execution. In (Yaqoob, 2017), the authors proposed a task migration mechanism, where suitable cloud agents are selected based on their higher task workload processing speed (CPU speed). In (Zhu, 2014), the authors proposed a real-time task scheduling scheme, in which the cloud server providing minimal energy consumption of the host mobile device is selected for migrated task processing. In (Shi, 2017), the authors proposed a task assignment scheme, in which an MU's task is migrated to a remote cloud server offering a lower predicted task workload processing delay. In (Gkatzikis, 2014), to minimize the transmission delay, the cloud server that is near the MU's task result download location is selected for task migration. Most of the existing studies on task migration considered either infrastructure-based cloud (remote cloud and cloudlet) or infrastructure-less cloud resources (mobile ad-hoc clouds) for task migration. Note, however, that these previous studies did not include different task properties (e.g., task arrival time and number, task workload, migrate task data size, and task deadline), bandwidth, and cloud agent properties (e.g., CPU speed, storage, location) for task assignment to suitable robot/cloud actors. Furthermore, they did not present any bandwidth allocation scheme serving both broadband and task migration traffic at the same time. Importantly, deployment scenarios considering different cloud agents/nearby mobile devices that may fail to satisfy given task offloading requirements have not explicitly studied in the past. There is also a lack of a multi-task resource scheduling mechanisms accounting for both user mobility and cloud agent failure avoidance. Finally, the performance evaluation of HART task

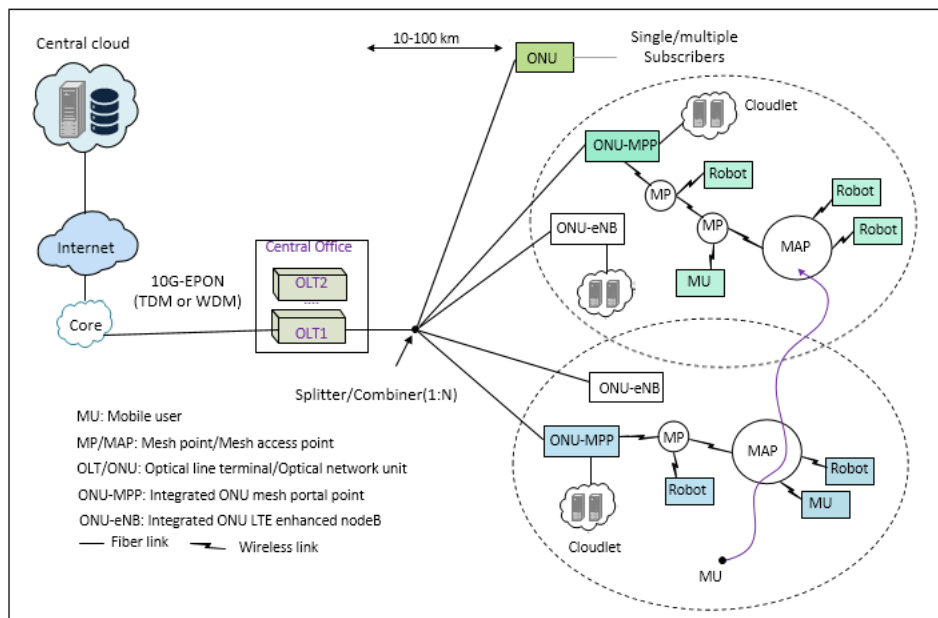
migration schemes in terms of both task execution time and energy consumption is missing as well. More specifically, no existing studies are available to calculate the various HART task completion delay components, including upstream and downstream task request transmission time, task workload processing, communication delay, offload packet buffering delay, and waiting delay at the cloud agents/robots.

PROPOSED HART TASK COORDINATION SCHEME

Network Architecture

Figure 1 depicts our proposed FiWi based Tactile Internet infrastructure used for enabling and coordinating HART-centric task coordination process, whereby humans, collaborative robots, and cloud agents actively participate in the joint task execution process. Our proposed FiWi based Tactile Internet infrastructure consists of a time division multiplexing (TDM) or wavelength division multiplexing (WDM) IEEE 802.3av 10 Gb/s Ethernet Passive Optical Network (10G-EPON) with a fiber backhaul range of 10-100 km between the central optical line terminal (OLT) and remote optical network units (ONUs). The OLT collocated with the central office serves three different subsets of ONUs, which are connected through a 1:N optical splitter/combiner at the remote node. One subset of ONUs may be equipped with a mesh portal point (MPP) and are henceforth referred to as ONU-MPPs, whereby mesh points (MPs) act as intermediate relay nodes between MPPs and mesh access points (MAPs). To provide 4G cellular services to MUs, another subset of ONUs is connected to an LTE enhanced nodeB (eNB) base station, giving rise to so-called ONU-eNBs. The central cloud servers are connected to the OLT via dedicated fiber links. In addition, local cloudlets are connected via dedicated fiber links to ONU-MPPs at the edge of our FiWi based Tactile Internet infrastructure in order to provide cloud services in close proximity

Figure 1. FiWi based Tactile Internet Infrastructures for HART Task Coordination

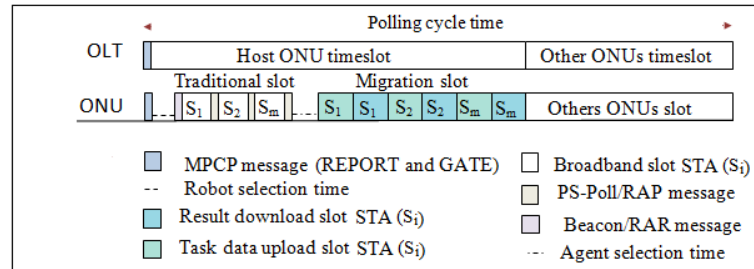


to nearby MUs and/or robots. For predicting the location of a given MU's task request transmission and task result reception, we adopt the widely used random waypoint (RWP) model (Deng, 2015).

Robot and Cloud Agent Selection for HART Task Execution

Our proposed latency sensitive HART task migration scheme, which selects a suitable robot for physical sub-task (e.g., image capturing at a task location) execution and a suitable agent (central cloud/local cloudlet/nearby robot) for cognitive sub-task (e.g., face detection from captured image) execution includes the following steps. Initially, an MU sends her full task request (both physical and cognitive sub-tasks) to the associated ONU (e.g., ONU-MPP/ONU-eNB). The task request message consists of information about the involved task type, task location, workload, deadline as well as the MU's initial location, speed, and task result download location. Subsequently, the ONU sends the task request to the OLT. Next, the OLT forwards the task request message to the host ONU (task coordinator) that serves at the respective task location. The task coordinator located at the host ONU receives the task request message and forwards it to all involved actor nodes (collaborative robots/robots and cloud agents). The actor nodes in turn send the task response message to the task coordinator node. Each actor's task response message contains information about the actor's task processing capability, busy time, and remaining energy. After receiving all actor nodes' task response messages, the task coordinator computes the predicted full task and physical sub-task execution time of all robots and cognitive sub-task execution time of all agents. In the non-migration scheme, the initially selected robot executes the full task. A suitable robot is selected according to the following criteria: satisfaction of task deadline, sufficient remaining energy, and minimum predicted task execution time among all robots.

Figure 2. Proposed Bandwidth Resource Management Scheme



Unified Bandwidth Assignment

As shown in Figure 2, in our proposed resource management scheme, the central OLT assigns bandwidth resources to ONUs (ONU-MPP/ONU-eNB) in the optical fiber backhaul via exchange of IEEE 802.3ah multipoint control protocol (MPCP) messages (REPORT and GATE). The ONU uses the REPORT message to inform the OLT about its next upstream transmission opportunity. The GATE message is used by the OLT to inform the ONU about its next upstream transmission timeslot. In the wireless front-end, ONUs assign bandwidth resources to associated wireless stations (STAs), which may be MUs or robots, via exchange of IEEE 802.11 messages, e.g., Beacon/PS-Poll message exchange between ONU-MPP and STAs for WLAN services and random-access preamble (RAP)/random access response (RAR) message exchange between ONU-eNB/STAs for cellular services (Hasan, 2013).

STAs use the PS-Poll/RAP message to inform ONUs(ONU-MPP/ONU-eNB) about their required bandwidth demands for the next polling cycle. Conversely, an ONU - upon reception of their own GATE message reception from the OLT - selects a suitable robot/cloud agent for each arriving task request, divides its timeslots into broadband and task migration sub-slots, and then assigns each time sub-slot to its associated STAs(MU/robot). Subsequently, the ONU sends a Beacon/RAR message to the STAs, which includes its assigned time sub-slot map for the next polling cycle. ONUs select suitable actors for each task request via exchange of DHCP messages (Discover-Offer-Request-Ack) between ONUs and involved actors(cloud/cloudlet/robot) (Rawadi, 2014). During their task migration upload sub-slot, STAs migrate their cognitive sub-task input data to the selected cloud agent for execution. In contrast, during their task migration download sub-slot, STAs receive the processed task result from the selected agent. For further technical details, the interested reader is referred to (Chowdhury, 2018).

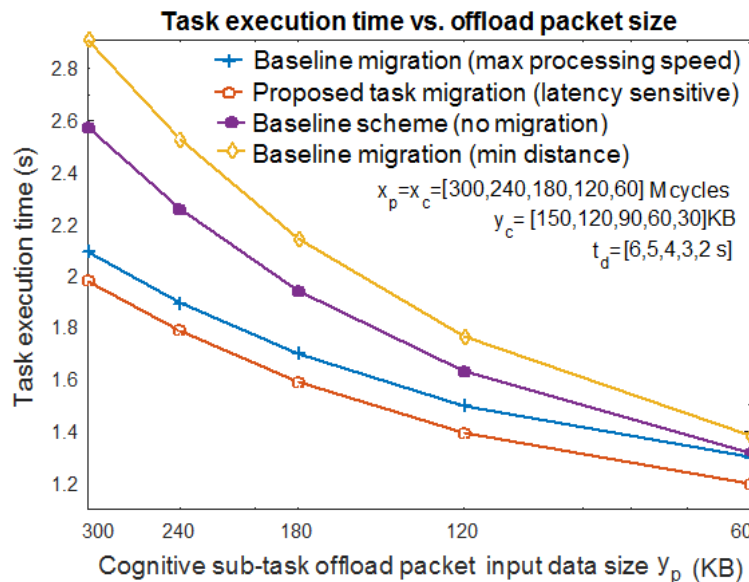
Predicted Task Execution Time Calculation

In this work, full task execution time includes the time duration between task assignment by task coordinator to suitable actors and task result reception by human mobile users. Hence, the predicted full task execution time of the non-migration scheme (only robot based) is calculated as follows:

$$t_{nm} = t_j^p + t_j^c = t_j^w + \frac{d_j^i}{q_m} + \frac{x_p}{q_p} + \frac{x_c}{q_p} + t_{j \rightarrow u}^r,$$

where t_j^p , t_j^c , t_j^w , d_j^i , q_m , q_p , x_p , x_c , and $t_{j \rightarrow u}^r$ denote the physical sub-task execution time of the robot, cognitive sub-task execution time of the robot, busy time of the robot, distance between the robot and the task location, the robot's moving speed, task workload processing speed, physical sub-task workload, cognitive sub-task workload, and task result transfer time (robot to MU), respectively.

Figure 3. Task Execution Time Evaluation



Power Consumption Value Calculation

In this work, STA's power consumption for full task execution includes the estimated power consumption of STA's (MU's and robots) for both physical and cognitive sub-task workload processing.

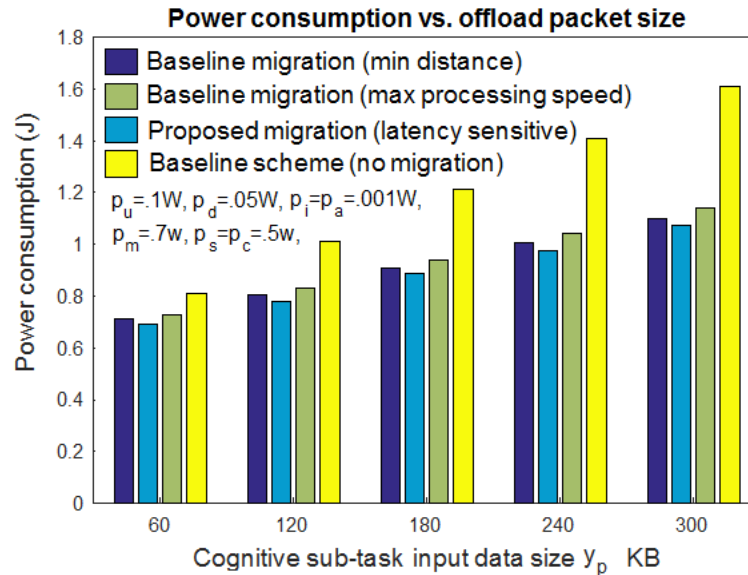
A wireless station's power consumption for full task execution in the non-task migration scheme is given by:

$$e_{nm} = e_j^p + e_j^c = p_i * t_j^w + p_m * \frac{d_j^i}{q_m} + p_s * \frac{x_p}{q_p} + p_c * \frac{x_c}{q_p} + p_d * t_{j \rightarrow u}^r,$$

where p_i , p_m , p_s , p_c , and p_d denotes the wireless station's average power consumption (per time unit) during the robot's busy time, moving time to task location, physical and cognitive sub-task workload processing time of the robot, and task result reception time, respectively. Hence, in our proposed full task execution with cognitive sub-migration scheme, both robot and agent are selected for executing the physical and cognitive sub-task, respectively. If more than one robot and agent satisfies both remaining energy and deadline criteria of the physical and cognitive sub-task, a suitable robot and cloud agent is selected based on the minimum predicted physical and cognitive sub-task execution time, respectively. Similarly, the full task execution time for task execution with migration (robot physical and agent cognitive sub-task execution) scheme is given by:

$$t_{wm} = t_j^p + t_a^c = t_j^w + \frac{d_j^i}{q_m} + \frac{x_p}{q_p} + t_{j \rightarrow a}^u + t_a^w + \frac{x_c}{q_a} + t_{a \rightarrow u}^d,$$

Figure 4. Power Consumption Evaluation



where t_j^p , t_a^c , t_j^w , t_a^w , q_a , $t_{j \rightarrow a}^u$, and $t_{a \rightarrow u}^d$ represent the physical sub-task execution time of the robot, cognitive sub-task execution time of the agent, busy time of robot and agent, agent cognitive sub-task workload processing speed (a=central cloud/cloudlet), cognitive sub-task upload (robot to agent), and task result download time (agent to MU), respectively. A wireless station's power consumption for the task execution with migration scheme is given by

$$e_{wm} = e_j^p + e_a^c = p_i * (t_j^w + t_a^w) + p_m * \frac{d_j^i}{q_m} + p_s * \frac{x_p}{q_p} + p_u * t_{j \rightarrow a}^u + p_a * \frac{x_c}{q_a} + p_d * t_{a \rightarrow u}^d,$$

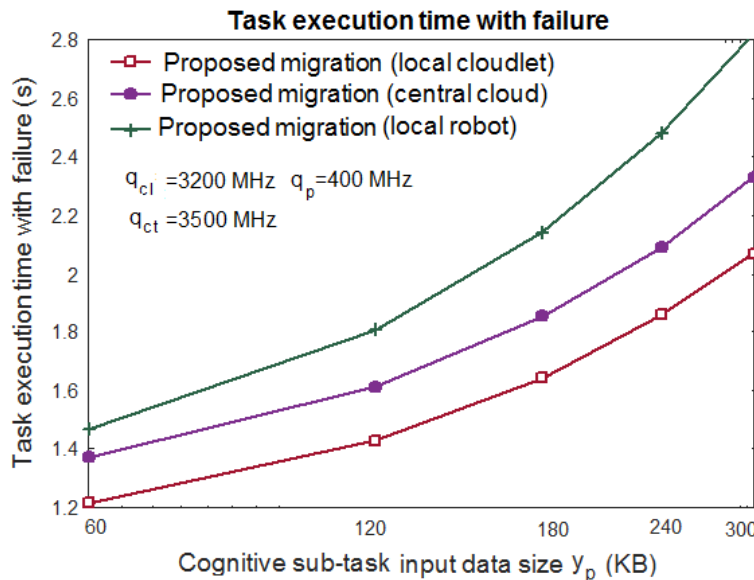
where p_r , p_m , p_s , p_a , p_u , and p_d denote the wireless station's average power consumption (per time unit) during the robot's busy time, moving time to task location, the robot's physical sub-task workload processing time, the agent's cognitive sub-task workload processing time, cognitive sub-task upload and task result reception time, respectively.

EVALUATIONS RESULTS AND FINDINGS

In the following, we investigate the performance of our proposed task migration scheme. For selecting a suitable task execution scheme, we compare the performance of the following four different schemes: (i) baseline scheme with no migration, (ii) proposed cognitive sub-task migration scheme (latency sensitive), (iii) baseline migration (maximum processing speed), and (iv) baseline migration (minimum distance).

In the no-migration scheme, only the initially selected robot (with minimum predicted execution delay among all robots) executes the full HART task that consists of both physical subtask (image capturing at task location) and cognitive subtask (face detection from captured image). Conversely, in the proposed migration scheme (latency sensitive), the initially selected robot executes the physical subtask

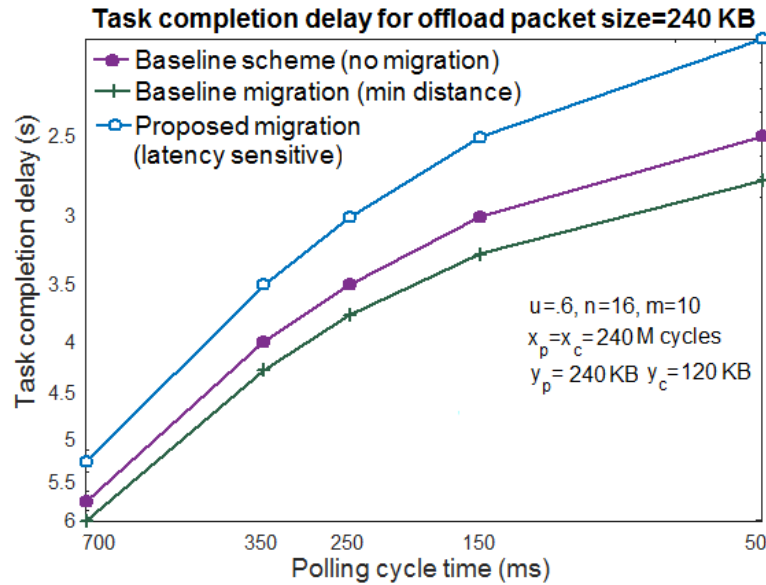
Figure 5. Task Execution Time with Failure Evaluation+--*



and migrates the cognitive subtask to a suitable cloud agent for execution. Hence, both the robot and cloud agent are selected based on their respective predicted lower physical and cognitive task execution time value, respectively. By contrast, in the baseline migration (maximum processing speed) scheme and baseline migration (minimum distance) scheme, actors (robot and cloud agent) are selected based on their maximum task processing speed and minimum distance, respectively. In case of multi-task arrivals, we assume that both actors (cloud agent/robot) and bandwidth resources are assigned to each task in a first-come-first-served (FCFS) manner. We also assume that each actor node is able to process only one sub-task (either physical or cognitive) at any given time. For evaluation, we vary the physical sub-task workload (x_p) and cognitive task workload (x_c), cognitive sub-task input (y_p), and output data size (y_c), and task deadline (t_d) values in compliance with (Zhang, 2013) and (Zhang, 2015). The legend in Figure 3 shows the default parameter settings of x_p , x_c , y_p , y_c , and t_d used in our evaluation.

Figures 3 and 4 highlight the task execution time and power consumption of the different schemes under consideration, respectively. Both figures show that the task execution time and power consumption grow rapidly for increasing cognitive sub-task input data size in all compared schemes. Both figures indicate that the minimum task execution time and power consumption are achieved in our proposed task migration scheme (latency sensitive). Among all compared schemes, the baseline migration (minimum distance) scheme exhibits the worst task execution latency performance, though it helps improve the power saving of the baseline scheme (non-migration scheme) due to the cognitive sub-task migration to another nearby actor at minimum distance.

Figure 6. Task Completion Delay Evaluation



We observe that both baseline migration (maximum processing speed) and baseline migration (minimum distance) don't outperform our proposed migration scheme (latency sensitive) in terms of task execution delay and power consumption. This is because in both schemes, either the maximum processing speed of the actor or the minimum distance is considered for selecting a suitable actor, respectively, but not both at the same time. Conversely, in our proposed scheme, suitable actors (robot/cloud agent) are selected based on the minimum predicted task execution time, including the robot/agent's physical/

cognitive sub-task workload processing delay, communication delay, and waiting delay at the robot/agent. For instance, for a typical cognitive sub-task input data size of 180 KB in Figure 3, our proposed migration scheme yields a 7.05%, 25.7%, and 18.55% higher task execution time efficiency than the baseline migration (maximum processing speed), baseline migration (minimum distance), and baseline scheme (no migration), respectively. For the same offload packet size of 180 KB in Figure 4, the achieved power consumption efficiency in our proposed scheme with regard to the baseline migration (maximum processing speed), baseline migration (minimum distance), and baseline schemes (no migration) is 7%, 4%, and 27% higher, respectively.

Figure 5 investigates the suitability of different actors for cognitive sub-task execution in case of initially selected cloud agent failure. The figure shows that compared with the central cloud (remote) agent and nearby local robot task migration schemes, the local cloudlet agent migration scheme experiences a lower task execution time delay due to its lower cognitive sub-task migration communication overhead and workload processing delay. Finally, Figure 6 illustrates the impact of the polling cycle time on the end-to-end task completion delay. Note that the end-to-end task completion delay is obtained by taking the time between task request transmission and task result reception by the MU. We observe that the task completion delay is small for lower polling cycle times and grows rapidly for increasing polling cycle time in all compared schemes. The figure also shows that for both longer and shorter polling cycle times, our proposed task migration scheme (latency sensitive) achieves a significantly lower end-to-end task completion delay than its counterparts and is thus more suitable. For instance, for a cognitive sub-task workload of 240 Mcycles, polling cycle time of 350 ms, total number of 16 ONUs, and FiWi traffic load of 0.6, the task completion delay efficiency of our proposed task migration (latency sensitive) scheme compared to the baseline (non-migration) and baseline migration (minimum distance) migration scheme is 13% and 18% higher, respectively.

FUTURE RESEARCH DIRECTIONS

Interesting future research studies that build on our proposed task migration scheme may unfold in the following possible directions. HART members can share their bandwidth, computation, and storage resources to stimulate beneficial cooperation, which in turn help cut down users' energy consumption and data usage cost. Hence, an efficient auction-based resource provisioning policy is mandatory to find the optimal price for human mobile users requesting task execution services. Another future research area of interest involves the use of HART task coordination for mixed-reality applications. For a cost-effective roll-out of mixed-reality applications, the development of adaptive real-time synchronization and computation techniques is another interesting direction of future research.

CONCLUSION

We investigated the challenges and benefits of HART-centric task execution over FiWi enhanced Tactile Internet infrastructures. To achieve higher task execution time and power consumption efficiency, we proposed a latency sensitive task migration scheme that selects a suitable collaborative robot for physical sub-task and a cloud agent for cognitive sub-task execution by taking a wide variety of task characteristics and actor (robot/cloud agent) resources into account. Further, we presented a resource allocation scheme that assigns bandwidth for transmission of broadband and task migration traffic over

FiWi network infrastructures. We presented comparison results of both HART task execution schemes with migration and without migration. Our presented results show that for a typical 120 KB (migrated) cognitive sub-task input data size, the achieved task execution time and power consumption efficiency in our proposed task migration scheme with regard to the baseline task execution (no migration) are improved by approximately 14% and 22.77%, respectively. Moreover, our obtained results highlight that when the initially selected agent node suffers from failure, local cloudlet agent cognitive task migration is more beneficial than remote cloud agent or local robot migration in terms of lower task execution time. Thus, our proposed task migration scheme is more suitable for lowering delay and power consumption inhuman-machine coactivity based on low-latency HART task execution.

REFERENCES

- Beyranvand, H., Levesque, M., Maier, M., Salehi, J. A., Verikoukis, C., & Tipper, D. (2017). Toward 5G: FiWi Enhanced LTE-A HetNets With Reliable Low-Latency Fiber Backhaul Sharing and WiFi Offloading. *IEEE/ACM Transactions on Networking*, 25(2), 690–707. doi:10.1109/TNET.2016.2599780
- Bradshaw, J., Dignum, V., Jonker, C., & Sierhuis, M. (2012). Human-Agent-Robot Teamwork. *IEEE Intelligent Systems*, 27(2), 8–13. doi:10.1109/MIS.2012.37
- Choudhury, B. (2009). Development of Optimal Strategies for Task Assignment in Multi-robot System. *Proc., IEEE IACC*, 1130–1135.
- Chowdhury, M., & Maier, M. (2017a). Local and Non-Local Human-to-Robot Task Allocation in Fiber-Wireless Multi-Robot Networks. *IEEE Systems Journal*, 12(3), 2250–2260. doi:10.1109/JSYST.2017.2661282
- Chowdhury, M., Steinbach, E., Kellerer, W., & Maier, M. (2018). Context-Aware Task Migration for HART-Centric Collaboration over FiWi Based Tactile Internet Infrastructures. *IEEE Transactions on Parallel and Distributed Systems*, 29(6), 1231–1246. doi:10.1109/TPDS.2018.2791406
- Daugherty, P., & Wilson, H. (2018). Human + Machine: Reimagining Work in the Age of AI. Harvard Business Review Press.
- Deng, S., Huang, L., Taheri, J., & Zomaya, A. Y. (2015). Offloading for Service Workflow in Mobile Cloud Computing. *IEEE Transactions on Parallel and Distributed Systems*, 26(2), 3317–3328. doi:10.1109/TPDS.2014.2381640
- Gkatzikis, L. (2014). Mobiles on Cloud Nine: Efficient Task Migration Policies for Cloud Computing Systems. *Proc., IEEE CloudNet*, 204–210. 10.1109/CloudNet.2014.6968993
- Gkatzikis, L., & Koutsopoulos, I. (2013). Migrate or Not? Exploiting Dynamic Task Migration in Mobile Cloud Computing Systems. *IEEE Wireless Communications Magazine*, 20(3), 24–32. doi:10.1109/MWC.2013.6549280
- Hasan, M., Hossain, E., & Niyato, D. (2013). Random Access for Machine-to-Machine Communication in LTE-Advanced Networks: Issues and Approaches. *IEEE Communications Magazine*, 51(6), 86–93. doi:10.1109/MCOM.2013.6525600

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Hoffman, R. R., Jonker, C., Riemsdijk, B., & Sierhuis, M. (2011). Beyond cooperative robotics: The central role of interdependence in coactive design. *IEEE Intelligent Systems*, 26(3), 81–88. doi:10.1109/MIS.2011.47

Maier, M. (2015). The Audacity of Fiber-Wireless (FiWi) Networks: Revisited for Clouds and Cloudlets. *China Communications. Feature Topic on Optical Interconnection Networks for Cloud Data Centers*, 12(8), 33–45.

Maier, M., Chowdhury, M., Rimal, B. P., & Van, D. P. (2016). The Tactile Internet: Vision, Recent Progress, and Open Challenges. *IEEE Communications Magazine*, 54(5), 138–145. doi:10.1109/MCOM.2016.7470948

Maier, M., Ebrahimzadeh, A., & Chowdhury, M. (2018). The Tactile Internet: Automation or Augmentation of the Human? *IEEE Access: Practical Innovations, Open Solutions*, 6, 41607–41618. doi:10.1109/ACCESS.2018.2861768

Mendonça, R. (2013). Efficient Distributed Algorithm of Dynamic Task Assignment for Swarm Robotics. *Computational Science and Its Applications, LNCS, Springer, 7971*, 500–510.

Rawadi, J. (2014). Providing Local Cloud Services to Mobile Devices with Inter-Cloudlet Communication. *Proc., IEEE MELECON*, 134–138. 10.1109/MELCON.2014.6820520

Shi, L., Zhang, Z., & Robertazzi, T. (2017). Energy-Aware Scheduling of Embarrassingly Parallel Jobs and Resource Allocation in Cloud. *IEEE Transactions on Parallel and Distributed Systems*, 28(6), 1607–1620. doi:10.1109/TPDS.2016.2625254

Trigui, S., Koubaa, A., Cheikhrouhou, O., Youssef, H., Bennaceur, H., Sriti, M.-F., & Javed, Y. (2014). A Distributed Market-Based Algorithm for Multi-Robot Assignment Problem. *Procedia Computer Science, Elsevier*, 32, 1108–1114. doi:10.1016/j.procs.2014.05.540

Yaqoob, I., Ahmed, E., Gani, A., Mokhtar, S., & Imran, M. (2017). Heterogeneity-Aware Task Allocation in Mobile Ad Hoc Cloud. *IEEE Access: Practical Innovations, Open Solutions*, 5, 1779–1795. doi:10.1109/ACCESS.2017.2669080

Zhang, H., Zhang, Q., & Du, X. (2015). Toward Vehicle-Assisted Cloud Computing for Smartphones. *IEEE Transactions on Vehicular Technology*, 64(12), 5610–5618. doi:10.1109/TVT.2015.2480004

Zhang, W. (2013). Energy-Efficient Scheduling Policy for Collaborative Execution in Mobile Cloud Computing. *Proc., IEEE INFOCOM*, 190–194. 10.1109/INFOCOM.2013.6566761

Zhu, X., Yang, L. T., Chen, H., Wang, J., Yin, S., & Liu, X. (2014). Real-Time Tasks Oriented Energy-Aware Scheduling in Virtualized Clouds. *IEEE Transactions on Cloud Computing*, 2(2), 168–180. doi:10.1109/TCC.2014.2310452

ADDITIONAL READING

Chowdhury, M., & Maier, M. (2017). Collaborative Computing For Advanced Tactile Internet Human-to-Robot (H2R) Communications in Integrated FiWi Multi-Robot Infrastructures. *IEEE Internet of Things Journal*, 4(6), 2142–2158. doi:10.1109/JIOT.2017.2761599

Fettweis, G. (2014). The Tactile Internet: Applications and Challenges. *IEEE Vehicular Technology Magazine*, 9(1), 64–70. doi:10.1109/MVT.2013.2295069

Liu, J., Shou, G., Liu, Y., Hu, Y., & Guo, Z. (2018). Performance Evaluation of Integrated Multi-access Edge Computing and Fiber-Wireless Access Networks. *IEEE Access: Practical Innovations, Open Solutions*, 6, 30269–30279. doi:10.1109/ACCESS.2018.2833619

KEY TERMS AND DEFINITIONS

Fiber-Wireless Enhanced Networks: Traditional FiWi access networks are based on integration of both optical fiber (Ethernet passive optical network or EPON) technologies (at the wired backhaul) and wireless (wireless local area network or WLAN) Ethernet technologies (at the wireless frond-end), which are then integrated with their cellular counterparts, namely, 4G Long Term Evolution Advanced (LTE-A), to give rise to FiWi enhanced LTE-A heterogeneous networks (HetNets).

Human-Agent-Robot Teamwork (HART): Human-agent-robot teamwork (HART) denotes the task and resource coordination framework for executing a requested task in a collaborative setting, in which human mobile device transfers the task request and receives the task result from collaborative machines (robot/cloud agent) after processing. Whereas, the robot and cloud agent process the physical (movement towards task location and image capturing using camera) and cognitive sub-task (face detection from captured image), respectively. Note that, in this work the robot uploads the physical sub-task output data size (input of cognitive sub-task) to cloud agent for processing. Hence, the cloud server/agent transfers the processed cognitive sub-task results to human MU's device after processing.

Mobile-Edge Computing: Mobile-edge computing (MEC) technology offers decentralized cloud computing services (e.g., computing storages) to mobile users' devices for executing their computation-intensive tasks/applications by setting up local cloud servers/cloudlet/fog servers at the network edge (WiFi/cellular base station) in close proximity to mobile users through a process named as computation offloading.

Predicted Task Execution Time: Predicted task execution time denotes the time interval between task request assignment by task coordinator to suitable actors (robot/cloud server) and task result reception by human MU's device.

Tactile Internet: The tactile internet is defined as a very low-latency communication system that ensures very low round-trip delay along with high availability, reliability, and security for real-time human-machine interaction-centric applications execution.

Task Migration: Contemporary mobile devices offload large amounts of computationally intensive tasks to resource-rich cloud servers for processing. Task migration broadens the scope of conventional computation task offloading by not only transferring the task from an MU to cloud servers/surrogates but also from one cloud server/surrogate to another one for execution.

Embedded Control System Design for Inverted Pendulum Type Mobile Robots Based on High-Level Petri Nets

Gen'ichi Yasuda

Nagasaki Institute of Applied Science, Japan

INTRODUCTION

In recent years, functionality and complexity of real-time controllers dramatically increased, beyond the complexity of hardware components. Due to the growing complexity of the software components in advanced multiprocessor-based robot controllers, including timing requirements of run-time elaboration of information from multiple external sensors, the need of early validation of the design of control software is increasing (David & Alla, 1994; Hruz & Zhou, 2007). Usually, control laws for non-linear physical systems like robotic vehicles and arms are specified in continuous time and allows for checking mainly qualitative properties like stability. At run time, these control laws are integrated and implemented as multi-tasks programs controlled by a real-time operating system on a single or multiprocessor target. The performance indices depend on the actual implementation, including sampling rates and computing latencies, besides the algorithm in use. Conventional research on real-time operating systems for complex robotic system control do not provide tools to analyze or synthesize such sampled control laws with respect to performance indices, measuring the impact of the organization of the control system on the overall controlled process. For the design and implementation of the overall software of real-time robotic control systems, while traditional techniques, such as finite state automaton, offer limited support for validating control algorithm design, the use of formal methods can increase the possibility of validating design.

This chapter presents a formal approach to designing software for sensor based non-linear motion control with a backstepping method (Khalil, 2002), focusing on the design of control software for non-holonomic two-wheeled inverted pendulum mobile robots (Nomura, 2009). Because the hardware of the control system exhibits subsystems with different dynamics, for example, feedforward paths to update some parameters and measurements to come from sensors of different kind running at different rates, control algorithms are decomposed into groups of cooperating tasks; where these tasks with different executing durations run in sequence or in parallel. Multi-task programs with multi-rate controllers permit modular programming and software reusability. The feature can be useful to optimize computing resources, because both onboard space and energy are strongly limited, especially for autonomous robots working in unstructured or critical environments.

BACKGROUND

To perform complex control software using multiple sensors and effectors, single or multiple computer control systems are essentially employed, where the overall management of a computer system is under

DOI: 10.4018/978-1-7998-3479-3.ch004

control of a multitasking operating system master program, which is referred to as a supervisor, monitor, or kernel. The master program is a collection of programs to schedule the jobs, to manage the hardware resources, and to provide a more hospitable interface to the user programs. So centralized operating systems have become so essential to efficient computer operation, viewed as inseparable from the hardware.

The simple Petri nets associate events with transitions and conditions with places. Places and transitions are linked by directed edges called arcs. Since the presence of a token in a place indicates that the corresponding condition is true, the marking describes the state of the Petri net in terms of conditions which are true and those which are false. When each of the input places becomes marked, i.e. has a token, the transition fires, i.e. the event occurs, and the firing removes a token from each of the input places, making the preconditions false, and deposits a token in each of the output places, making the postconditions true.

The Petri net is gaining increasing importance for the discrete event modeling and analysis of robotic control systems, since it offers a convenient way of expressing system behavior which is both parallel and asynchronous, and therefore distributed (Wang & Sarides, 1993; Yasuda, 2014; 2015a). In addition to the precedence constraints among computing modules, including sequences of modules and the repetition of certain modules, looser couplings associated with shared resources, such as memory and devices, can also be directly expressed. Besides modeling capabilities, the Petri net can be analyzed in a formal way to obtain information about the dynamic behavior of the modeled system (Du et al., 2013).

DISTRIBUTED CONTROLLER DESIGN BASED ON BACKSTEPPING APPROACH

A wheeled inverted pendulum mobile robot is a self-balancing, traveling vehicle with two wheels attached on the sides of its body. Two motors are attached between the wheels and the body to control traveling speed and direction, besides its own posture. Both the traveling motion and the self-balancing are achieved by independent actuators providing the necessary torques to the wheels (Fierro & Lewis, 1997; Nomura, 2009; Takei et al., 2009). The complete dynamics of the traveling motion consists of the steering kinematics and the vehicle dynamics. The nonholonomic constraint states that the vehicle can only move in the direction normal to the axis of the driving wheels, satisfying the conditions of pure rolling and non-slipping. For traveling control, divided into tracking a reference trajectory, following a path, and point stabilization, the software generates robot trajectories based on on-line elaboration of sensors information and manages kinematic inversion. According to a backstepping approach, the main idea behind nonlinear feedback control for solving these problems is to define velocity control inputs based on the perfect velocity tracking assumption. According to a backstepping approach, nonlinear feedback control for solving these problems velocity control inputs are defined based on the perfect velocity tracking assumption. So, velocity control inputs should be converted into torques, taking into account the actual vehicle dynamics.

For balancing control, the posture angle x_1 and its derivative $x_2 = \dot{x}_1$ are used as the state variables, where the posture angle is 0 when the center of gravity of the robot is right above the contact point of a wheel to the ground. The reference values of the posture angle and the angular velocity are zero. The state-space equations of balancing control are generally represented as follows:

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{1}{g_1(x_1)}(u - h(x_1, x_2))\end{aligned}\quad (1)$$

As a first step of backstepping approach, the virtual control law is defined such that for the posture angle error $z_1 = x_1^{ref} - x_1$, $\lim_{t \rightarrow \infty} z_1(t) = 0$, as follows:

$$\alpha = k_1 z_1 + k_2 \int z_1(\tau) d\tau \quad (2)$$

where k_1 and k_2 are positive constants.

Then, an input torque u is designed such that for the error $z_2 = \alpha - x_2$, $\lim_{t \rightarrow \infty} z_2(t) = 0$. The first and second Lyapunov functions are defined as:

$$V_1 = \frac{1}{2} z_1^2 + \frac{k_2}{2} \left(\int z_1(\tau) d\tau \right)^2 \quad (3)$$

$$V_2 = V_1 + \frac{1}{2} z_2^2 \quad (4)$$

Differentiating V_2 as follows:

$$\dot{V}_2 = \dot{V}_1 + z_2 \dot{z}_2 = -k_1 z_1^2 + z_1 z_2 + z_2 \dot{z}_2 \quad (5)$$

For $\dot{V}_2 < 0$, $\dot{z}_2$ is chosen as follows:

$$\dot{z}_2 = -k_3 z_2 - z_1 \quad (6)$$

where k_3 is a positive constant.

Consequently, the input torque is determined as:

$$u = G(z_1, \int z_1(\tau) d\tau, z_2) \quad (7)$$

Since $\dot{V}_2 \leq 0$ is derived, it is ensured that the tracking error converges on zero.

Figure 1 shows the overall structure of software modules for generating motion control of two-wheeled inverted pendulum mobile robot. The motion controller mainly consists of the following cooperative computing modules: path planning, self-localization, motion planning, motion monitoring, balancing and traveling control, and exteroceptive sensing (wheel rotation and angular velocity sensing, and body posture angle and angular velocity sensing). The software module for balancing and traveling control obtains the wheel rotational angles and their angular velocities and the body posture angle and angular velocity from the respective sensing modules at each sampling time and controls the torque of the two

motors. The gyro sensor output voltage is proportional to the body posture angular velocity and the posture angle is obtained by an accumulation of the angular velocity, where scale factors of offset or drift and sensitivity of the sensor are calibrated. The self-localization module estimates the robot position based on odometry through the traveling distance and orientation of the body obtained from the wheel rotation and angular velocity module, which are calculated by counting the number of encoder pulses.

The path planning module computes nominal trajectories in terms of positional setpoints or subgoals in the operational space and passes to the motion planning module, which corrects them according to balancing control, performing velocity calculation by kinematic inversion, both with a repetition interval of 10 ms. The motion planning module generates a plan of motion, including velocity pattern and moving direction, to the next subgoal. The traveling control module communicates with the motion planning module and executes micro-interpolation and position control in joint space of the vehicle at a sampling interval of 1 ms. The robot travels from a subgoal to the next subgoal along a given path on the map while keeping its posture. At each subgoal, the completion signal is sent to the monitoring module, and the estimated position is adjusted using laser range sensor data and the map. Figure 2 shows a view of the wheeled inverted pendulum mobile robot using LEGO Mindstorms NXT.

Figure 1. Software structure of motion control of two-wheeled inverted pendulum mobile robot

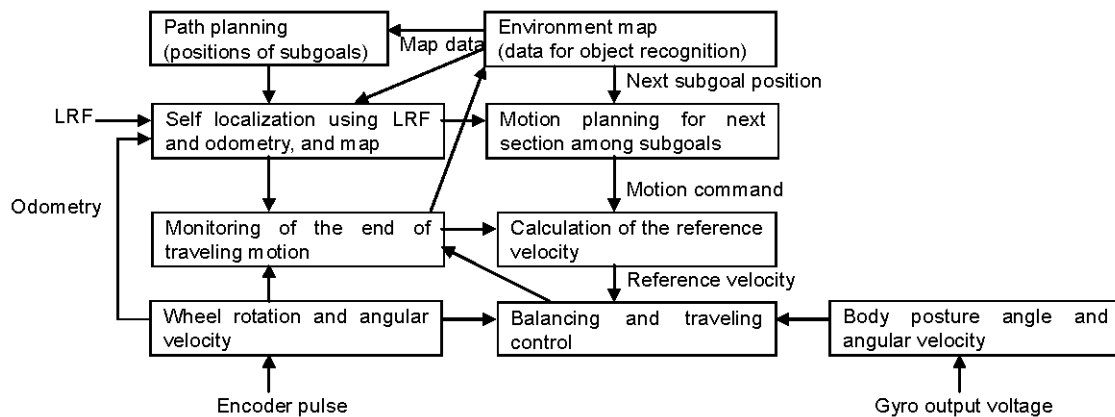


Figure 2. View of experimental mobile robot



Additionally, some sensor modules are provided to handle the acquisition of exteroceptive gyro sensor data and encoder pulses from the wheels. The overall control system has been configured to introduce exteroceptive sensor-based control capabilities as an evolution of the current industrial controller. The control unit is based on the VME-bus and includes a multiprocessor board equipped with Motorola 68020/68882 CPU's for path and motion planning, and a DSP (Texas 320C25) for balancing control and traveling control. During control of physical processes, data always flow from sensors to actuators along several control paths. Since feedforward paths and parameter adjustments are typically executed more slowly (less frequently) than feedback error computations, multi-rate sampling can optimize computing power.

DISCRETE EVENT MODELING AND CONTROL NET REPRESENTATION

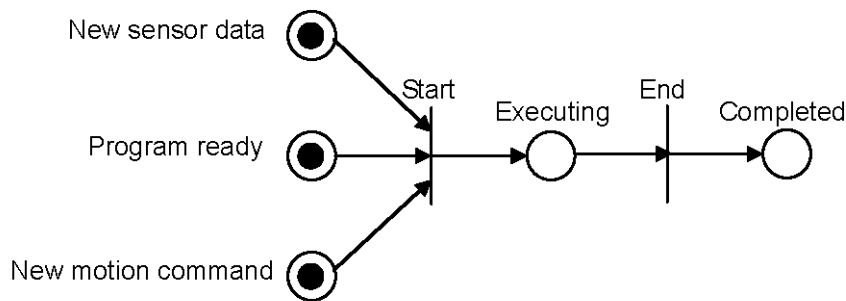
The control system development methodology consists of three phases to assist Petri net based implementation of the control software. In the first phase, the robotic task and environment requirements are specified to identify functionalities and attributes of the hardware and software architecture needed to complete the task. In the second phase, the net model for system control is designed as follows. First, the net model of the task or recipe specifications is constructed using the task-oriented approach. Second, the net model of resource specifications corresponding to each task specification is provided using the mutual exclusion concept. Then, the task and resource net models are integrated to yield the basic net model for system control. In the third phase, by analyzing and refining the basic net model, a deadlock-free, safe, and reversible system net model is obtained. For non-centralized distributed control, the system net model may be decomposed with respect to resources, into net models of communicating intelligent subsystems, which are autonomously operated using corresponding resources or devices.

A customized Unified Modeling Language (UML) framework (Eshuis & Wieringa, 2004) is used to describe every component of the control system, including class attributes, methods, and their relations of inheritance, composition and association providing a structural view of the software with the hardware locations of the components. Based on the above class diagrams, the interaction sequence and the messages passed among classes are described using UML sequence, collaboration, and use-case diagrams. The statecharts are generated to provide a global view of the inherent states of the software components behavior. Then, the dynamics of the UML schematics is translated to Petri nets. Nowadays, the state diagrams or statecharts are adopted to extract the discrete event system dynamics of the UML framework design. Statecharts are designed as state machines and built by initial and final pseudostates, and transitions having one input state and one output state. Each transition arrow has a label with three optional parts: event, condition, activity; if the current state is the specified source state and the appropriate event occurs with the specified condition, the transition is fired and the specified action is taken. The translation from a UML statechart to a Petri net is simple such that each statechart start is converted into a Petri net place and each statechart transition into a Petri net transition with corresponding labels and tokens. Finite state machines are equivalently represented by a Petri net subclass, where a place having two or more output transitions is referred to as a conflict, decision or choice and the representation of the synchronization of parallel activities does not be allowed.

The execution of an active program with finite duration can be modeled with two events; start transition that represents the beginning of the computing action, end transition that represent the termination of the action, and place that models the action in progress. Figure 3 shows a Petri net model of computing module in motion control software. Based on a bottom-up approach, the first terms to be designed are

the device-level tasks or closed loop control laws encapsulated in a reactive logical behavior. The main features of the device-level control software are modeled using extended notations based on the ordinary Petri nets, where events are modeled by transitions as instantaneous actions occurring in a system. States can either represent a passive condition, such as waiting for input data, or the execution of an active program, for path planning, object recognition, or control computation. In repetitive execution, control data can be associated as integer iteration number with tokens in order to control transition firing by means of predicates. As a simple representation of interaction with the environment, a firing transition may trap sensory data and send output data to actuators. The transitions behave as agents for communication with the environment by use of the pre- and postprocessors of a firing transition; an enabled transition can call a preprocessor before it starts firing and after firing it can call a postprocessor. The preprocessor also sends signal to the actuator to turn on its effector.

Figure 3. Petri net model of computing module in motion control software



A task is the maximum granularity considered by the control systems engineer at the device control level, characterizing continuous time closed loop control laws, with their temporal features related to the management of associated events. The continuous time specification is translated into a description by adding temporal properties, such as discretization of time, durations of computations, communication and synchronization between the involved processes. Consequently, each task is defined in terms of communicating real-time computing modules, which each implement an elementary part of the control law. Most of the modules are periodically executed, generating interrupts needed to activate control tasks each sampling time with a real-time clock, to perform the calculations involved in the computation of the control algorithm, for example, state observer, estimation, sensor output compensation, kinematic inversion, calculation of virtual control, and calculation of control torque. Furthermore, other modules monitor conditions, which are classified into preconditions, post conditions, and exceptions like reaching a task space limit, and may trigger events participating in the reactive management of the task. A typical task is sequentially composed of the following modules: reading an input port, performing a calculation, and writing to an output port, and may be modeled by a condition/event net with three transitions for start and end of the actions and a transition generating interrupts with different periodic events by a real-time clock which is also modeled by a Petri net. By associating durations with transitions or places, a timed Petri net can be obtained. Considering the duration of calculation of the transitions, reading and writing can be assumed to be instantaneous events.

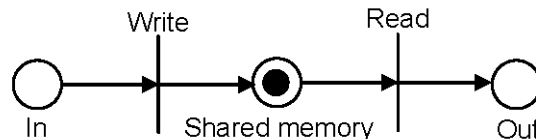
DESIGN OF CONTROL AND COMMUNICATION NETS

In case that each module is running freely or independently and the communication does not add synchronization, it can lead to an unexpected ordering of modules at run time by the operating system without any supervisor, which is similar to the unordered instinctive situations at the unconscious level within the human brain. So, to translate control laws into a system net model of task, a synchronization between modules scheme should be adopted for ensuring that the system net model is deadlock free and minimizing computing latencies to improve control performance.

In modeling control software with periodic control and communication laws, a basic property to be ensured is structural deadlock freedom, which means that given any natural initial marking, deadlock cannot occur. This property has corresponding definitions in the Petri net. A Petri net is live with respect to an initial marking if, from any marking in the set of markings reachable from the initial marking, there exists for each transition a legal firing sequence leading to a marking in which that transition is enabled. Because deadlock arises when no transition can be longer enabled, liveness implies deadlock freedom. Partially synchronizing some computing modules can improve the practical performance of the implementation. However, using synchronizations may lead to deadlocks or temporal inconsistencies. The consequences of introducing such synchronization should be examined in terms of structural and temporal problems using Petri net modeling and analysis. Some guidelines about how to add such synchronization are given to design deadlock-free and efficient implementations of real-time periodic control laws (Liu et al., 2013). In case each place has just one input transition and one output transition, that is, each associated condition may only become true in one way and can have one consequence, the task behavior is deterministic and the resulting Petri net is called a marked graph. For a marked graph the token count in a directed circuit is invariant under any firing (Simon et al., 1998). Using the marked graph, parallel construct of modules and their synchronization can be modeled with two transitions (events): transition that enables two parallel modules and transition that fires when both modules are completed, where both start transitions fire with no delay and end transitions are delayed by the duration of the modeled actions of the modules.

Generally, shared memories are used to exchanging data among modules, holding data produced by some modules and consumed by other modules. A shared memory is modeled with a place marked with a token that stores the shared data. Shared memories with capacity 1 and no access control, which store only one value that can be written or read without limitations, can be used for storing sampled data to be used for some modules, because only the most recent value is important regardless of the use of older values (Silva et al., 2014). Figure 4 shows a net model of shared memory.

Figure 4. Petri net model of shared memory



For typical motion commands, the planning module and the sensing module is activated synchronously at the established sampling frequencies as follows. The planning module computes nominal trajectories in terms of operational space position setpoints and passes them to the sensing module, which corrects

them according to measured data and control algorithms and then returns the corrected setpoints to the planning module. Setpoints and corrected setpoints are asynchronously communicated from the planning module to the sensing module and in reverse, respectively, through the shared memory. Trajectory correction is computed and applied to the reference trajectory generated by the planning module. Path profiles are handed over to the planning module. On the other hand, for sensor-based motion commands, the sensor module detects runtime the next goal through a vision or distance sensors and passes a sequence of motion commands or subgoals to the planning module. The communication between the two modules should be asynchronous because the planning module has to complete the execution of the current motion command before it can handle the next one.

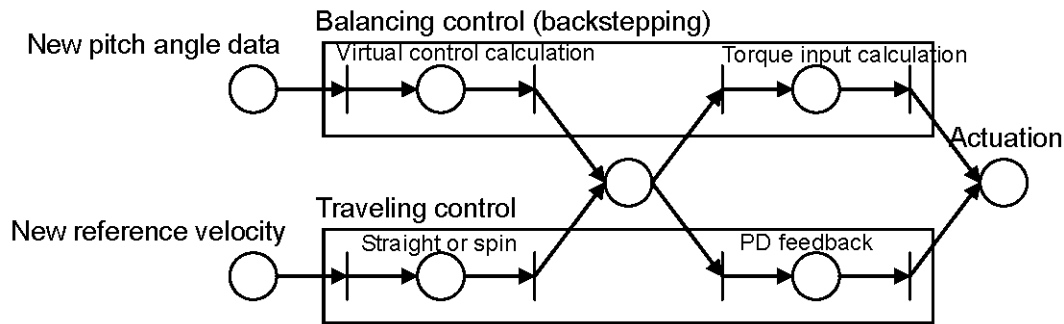
The procedure to merge two Petri nets to generate a composed module or task using shared memory is as follows. For corresponding transitions between two Petri nets, a communication place labeled as “Signal sent” or “Buffer” is built, and an output arc from the transition labeled as “Send (Write)” to the place and an input arc from the place to the transition labeled as “Receive (Read)” are added. If there is a transition that corresponds to a transition already connected to a communication place, then an input or output arc between the transition and the communication place is added. Because initial marking of the composed Petri net must be referenced to the Petri nets, numerical labels of places and transitions are reorganized to create a joined marking and transition enumeration for the composed Petri net.

Besides shared memory, different existing components, such as flags, toggles, which satisfy the modeling requirement of the multitasking control applications, can be modeled to link pairs of modules via their ports. As a more practical method of communication between the sender and receiver modules, a flag records a Boolean value with two places, either of which holds a token indicating the current state of the flag, and it can be set to either of its values independently from its previous value. The modules communicate synchronously through a set of shared flags to notify the states of execution of control tasks. For example, when the path planning module is completed, the corresponding flag is set. Then the transitions for the sensing module can be enabled and the module will start. The value of a flag can be tested checking for the presence of a token in the places. Since transitions need to remove a token to test a flag state, they are also required to put the token back so that the state of the flag will be preserved.

Communication between planning and control modules can be asynchronous or synchronous depending motion commands, and the protocol should be modified so that the last or the next available one is the best data for closed-loop control without loss of data. In synchronous communication, each module acts synchronously with the other; the first module to reach the rendez-vous is blocked until the second one is ready. In asynchronous communication, while the sender is running freely and posts messages on its output ports at each period, the receiver either receives the message if a new one is available or is blocked until the next message is posted. Symmetrically, the receiver runs freely, the sender is blocked up to the next request except if a new message was posted since the last receiving.

Although designing a system model of a task using Petri nets promote a modular approach (Fleury et al., 1994; Music & Matko, 1999), mapping each net model to a thread at run time increases operating system overhead due to numerous context switches and inter-process communication under the centralized supervising computer control. Therefore, assigning a group of modules which are sequentially executed to a thread by adding a certain amount of synchronization is desirable, where communication between groups or threads does not add synchronization constraints, so that the synchronization skeleton of the task is not changed. The device-level tasks are logically and temporally composed to build more complex tasks, then the higher-level software is designed in terms of a set of robotic tasks to be executed under a real-time operating system. Figure 5 shows a detailed net model of backstepping controller using shared memory and flags.

Figure 5. Petri net model of backstepping controller



IMPLEMENTATION ISSUES ON A MULTITASK PROCESSING ARCHITECTURE

For periodic and multi-rate modules distributed over a multiprocessor target architecture, various message passing mechanisms over type ports have been implemented. The structure of the periodic modules is provided by separating calculations, communications and calls to the underlying operating system. The temporal attributes of a module for calculation are duration of execution, activation period, input and output ports and their communication protocols, priority for scheduling at run time, and assigned processor in case of a multiprocessors target. The non-periodic reactive behavior of the task is handled by a special module which may be awakened by events coming from the task itself or others; these events represent the external view of the task and are used to compose tasks into more complex tasks.

As a special task for navigation, the console is modeled with the transition, and its input and output places. The transition represents an instruction loader, by popping a command from a queue of commands according to its time-stamp in the input place and putting it in the output place which is connected to the sensor module for execution of sensor-based control algorithms. The abstract data type of the queue is defined with the usual operations such as popping which extracts the top element from the queue, testing which is true if the queue is empty, etc. The place which represents the interface of the sensor contains the current value of the sensor. The input transition simulates the sensor, by updating the value of the token in the place, popped from the queue in the input place of the transition.

Application programs for Petri net based modeling and simulation of discrete event dynamic systems have been developed and used as a real-time control simulator. The Petri net simulator is realized as a toolbox for the MATLAB platform, so that the general purpose MATLAB tools for simulation of communication actions such as SimEvents and SIMULINK can be used in the Petri net models. The MATLAB environment with RWTH-Mindstorms NXT Toolbox allows control of LEGO Mindstorms NXT through Bluetooth wireless connection Data Acquisition Toolbox for receiving and sending input and output signals from/to external sensors and effectors, communicating with USB hardware. The execution of the simulator keeps periodically scanning; in each scan a Petri net is executed in the similar manner as a multitasking real-time operating system. The simulator checks for any enabled transitions. If there is any enabled transition it is placed into a queue-like data structure, then the respective transition definition file is run to make sure whether it can actually fire, satisfying all the guard conditions defined in the file including reserving resources for use by it. If an enabled transition can fire, then it is moved to another data structure. At the same time, input tokens for this transition are consumed from the respective input places. The simulator also checks whether any firing transition is completing. If there is any transition that is completing, then it is removing from the structure and the output tokens

are deposited into the output places of the transition. For optimal management of system resources, the following functions are provided: requesting and releasing resources, declaring and changing priorities of different transitions, and reporting resource usage idle time.

CONCLUSION

Because the backstepping control design can promote distributed implementation of the complex robotic control software, this chapter has described a unified and formal approach of modular, cooperative control software synthesis for complex mobile robot hardware using extended high-level Petri nets. The control software can be executed by simulating the overall Petri net models, such that the simulated behavior is the same as the actual behavior of the controlled objects or sensors and effectors. Consequently, the modeling and control of robotic motion can be executed in real time on a general computing architecture such as a client server system connected on an Ethernet network, where the system controller is the server and the local controllers are the clients. In contrast to conventional multitasking operating systems, the described Petri net based coordination method proposes a non-centralized control scheme without supervisor, monitor, or kernel. For robust-adaptive controller based on neural networks and more complex command tasks with real-time trajectory tracking including obstacle avoidance, the net models can be easily implemented, but it should be more efficiently executed on a high-performance multiprocessor architecture.

FUTURE RESEARCH DIRECTIONS

Due to the use of hierarchical, modular Petri nets, the composition/decomposition strategies for obtaining subnets or agents may be adapted to configure a bio-inspired autonomous intelligent hierarchical agent-based control system, according to the number of processors, the similarity structure of the subnets or the geographical distribution of the subsystems (Yasuda, 2015b; 2016). Each controller can be passively and asynchronously activated as a high-level agent in response to corresponding external and/or internal conditions and independently and concurrently executed. hierarchical agent-based control system. In a hierarchical agent-based control system, some local events, such as program start, end, and failure, generated by lower-level agents are monitored as global events and consumed by the higher-level agents, which globally induce lower-level agents to activate corresponding collective behaviors, like neural correlates of consciousness in human brains.

REFERENCES

- David, R., & Alla, H. (1994). Petri nets for modeling of dynamic systems. *Automatica*, 30(2), 175–202. doi:10.1016/0005-1098(94)90024-8
- Du, Y., Jiang, C. J., & Zhou, M. C. (2013). Modeling and analysis of real-time cooperative systems using Petri nets. *IEEE Transactions on Systems, Man, and Cybernetics. Systems, Part A Systems and Humans*, 37(5), 643–654. doi:10.1109/TSMCA.2007.902622

- Eshuis, R., & Wieringa, R. (2004). Tool support for verifying UML activity diagrams. *IEEE Transactions on Software Engineering*, 7(30), 437–447. doi:10.1109/TSE.2004.33
- Fierro, R., & Lewis, F. L. (1997). Control of a nonholonomic mobile robot: Backstepping kinematics into dynamics. *Journal of Robotic Systems*, 14(3), 149–163. doi:10.1002/(SICI)1097-4563(199703)14:3<149::AID-ROB1>3.0.CO;2-R
- Fleury, S., Herbb, M., & Chatila, R. (1994). Design of a modular architecture for autonomous robots. *IEEE International Conference on Robotics and Automation*. 10.1109/ROBOT.1994.351031
- Hruz, B., & Zhou, M. C. (2007). *Modeling and Control of Discrete Event Dynamic Systems*. Springer Verlag.
- Khalil, H. K. (2002). *Nonlinear Systems*. Prentice Hall.
- Liu, G. J., Jiang, C. J., Zhou, M. C., & Xiong, P. C. (2013). Interactive Petri Nets. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, 43(2), 291–302. doi:10.1109/TSMCA.2012.2204741
- Music, G., & Matko, D. (1999). Petri net based control of a modular production system. *IEEE International Symposium on Industrial Electronics*, 3, 1383–1388. 10.1109/ISIE.1999.796909
- Nomura, T., Kitsuka, Y., Suemitsu, H., & Matsuo, T. (2009). Adaptive backstepping control for a two-wheeled autonomous robot. *ICROS-SICE International Joint Conference 2009*, 4687–4692.
- Silva, E., Campos-Rebelo, R., Hirashima, T., Moutinho, F., Malo, P., Costa, A., & Gomes, L. (2014). Communication support for Petri nets based distributed controllers. *Proceedings of 2014 IEEE International Symposium on Industrial Electronics*, 1111–1116. 10.1109/ISIE.2014.6864769
- Simon, D., Castaneda, E. C., & Freedman, P. (1998). Design and analysis of synchronization for real-time closed-loop control in robotics. *IEEE Transactions on Control Systems Technology*, 6(4), 445–461. doi:10.1109/87.701336
- Takei, T., Imamura, R., & Yuta, S. (2009). Baggage transportation and navigation by a wheeled inverted pendulum mobile robot. *IEEE Transactions on Industrial Electronics*, 56(10), 3985–3994. doi:10.1109/TIE.2009.2027252
- Wang, F. Y., & Sarides, G. N. (1993). Task translation and integration specification in intelligent machines. *IEEE Transactions on Robotics and Automation*, RA-9(3), 257–271. doi:10.1109/70.240195
- Yasuda, G. (2013). Design of a sensor-based distributed behavioral control architecture for autonomous mobile robots with Petri Nets. *Proceedings of SICE Annual Conference 2013*, 1716–1721.
- Yasuda, G. (2015a). Distributed autonomous control of mobile robots based on hierarchical task decomposition and mutual cooperation by communication. *Proceeding of the 2015 JSME Conference on Robotics and Mechatronics*.
- Yasuda, G. (2015b). Design of distributed robot system controllers based on hierarchical modular structure of robotic tasks. *Proceeding of the 58th Joint Automatic Control Conference of Japan*.
- Yasuda, G. (2016). Control program design of networked distributed mobile robot systems using discrete event net based inter-task synchronous communication. *Proceeding of the 2016 JSME Conference on Robotics and Mechatronics*.

ADDITIONAL READING

Grehan, R., Moote, R., & Cyliax, I. (1998). *Real-Time Programming: A Guide to 32-bit Embedded Development*. Addison Wesley.

KEY TERMS AND DEFINITIONS

Backstepping: A technique for designing stabilizing controls for a special class of nonlinear dynamical systems, which are built from subsystems that radiate out from an irreducible subsystem that can be stabilized using some other method. The design process is a recursive structure, starting at the known-stable system and backing out new controllers that progressively stabilize each outer subsystem. In each step a Lyapunov function and a virtual control law are derived for each subsystem.

Deadlock: A situation that occurs when tasks prevent each other from running because of a cyclic locking and requesting of resources. The task may have to wait to acquire any shared resource.

Multitasking: Computer's capability to perform more than one task at the same time, switching from task to task quickly and seamlessly, although it is processing only one instruction at a time.

Shared Memory: A region of memory shared by multiple processes, provided for interprocess communication; each process that accesses shared memory sees the memory as though it were part of that process's private memory space.

Synchronization: In a multitasking system, where tasks interact directly with each other or indirectly through shared resources, coordination of their interaction to prevent undesirable race conditions.

Task: A task is a generic term that refers to an activity carried out by software that occurs independently from other activities in the system. Tasks that are separate executing programs are called processes.

Trajectory Tracking: Given a reference trajectory in the plane, a navigation technique that computes the torque input such that the trajectory error tends to zero.

Wheeled Inverted Pendulum Mobile Robot: A vehicle that controls its own posture, traveling speed, and direction with two wheels attached on the sides of its body.

Semantic and Formal Representation of Cognitive Models for the Metacognitive Architecture CARINA

Alba J. Jerónimo

 <https://orcid.org/0000-0001-9397-555X>

Universidad de Córdoba, Colombia

María P. Barrera

 <https://orcid.org/0000-0003-0652-5927>

Universidad de Córdoba, Colombia

Manuel F. Caro

 <https://orcid.org/0000-0001-6594-2187>

Universidad de Córdoba, Colombia

Adán A. Gómez

 <https://orcid.org/0000-0002-7676-8492>

Universidad de Córdoba, Colombia

INTRODUCTION

A cognitive model is a computational model of internal information processing mechanisms of the brain for the purposes of comprehension and prediction. Cognitive models are developed to study different aspects of cognition; attention and multitasking, judgment and choice in decision-making and skill acquisition in dynamic situations. CARINA metacognitive architecture runs cognitive models. However, CARINA does not currently have mechanisms to store and learn from cognitive models executed in the past. Semantic Knowledge Representation is a field of study which concentrates on using formal symbols to a collection of propositions, objects, object properties and relations among objects. In CARINA Beliefs are a form of represent the Semantic Knowledge, which are elements about facts, concepts or objects. The aim of this chapter is to formally describe a CARINA-based cognitive model through of Denotational Mathematics and to represent these models using a technique of Semantic Knowledge Representation called Beliefs. A CARINA-based cognitive model contains a problem, goals, mental states, beliefs, actions and reasoning rules. All the knowledge received by CARINA is stored in the semantic memory in the form of Beliefs. Thus, a cognitive model represented through beliefs will be ready to be stored in semantic memory of the Metacognitive Architecture CARINA. Finally, we present an illustrative example that describes a cognitive model of a syntactic analyzer for opening questions using beliefs, a structured natural language notation named NGOMS-L and a domain-specific visual language called M++.

Semantic Knowledge Representation is a way of modeling and introduce knowledge using tools that represent notions or concepts (Rajman & Chappelier, 2010) as well as formal symbols in a collection of propositions. Semantic Knowledge Representation is the area of Artificial Intelligence concerned how

DOI: 10.4018/978-1-7998-3479-3.ch005

knowledge can be symbolically manipulated and represented in an automated way by reasoning programs (Levesque, 2004). A cognitive model is a computational model of some internal information processing mechanisms of the brain used in the area of Artificial Intelligence for the purposes of comprehension and prediction (Levesque, 2004). Also, a cognitive model is a theoretically grounded and empirically guided specification of the mental representations and processes involved in a given cognitive function (Zacarias, Magalhaes, Caetano, Pinto, & Tribolet, 2008). Cognitive architectures have been used to create cognitive models of a variety of intelligent systems (Langley, Laird, & Rogers, 2009). A cognitive model of a task constructed in a cognitive architecture is runnable and produces a sequence of behaviors (Cox, Oates, & Perlis, 2011). Cognitive architectures have been used to create cognitive models of a variety of intelligent systems (Forstmann & Wagenmakers, 2015). A cognitive architecture is a general-purpose control system inspired by scientific theories developed to explain cognition in animals and humans (Langley et al., 2009).

A Metacognitive Architecture is a cognitive structure which is able to reasoning about itself (Cox et al., 2011). CARINA is a metacognitive architecture for artificial intelligent agents (Caro et al., 2018). CARINA is derived from the MISM Metacognitive Metamodel (Caro, Josyula, Cox, & Jiménez, 2014). When the reasoner executes a cognitive model in CARINA it is necessary that the cognitive model that was executed be stored in the Semantic Memory in form of beliefs for to be subsequently retrieved. Several researches have presented different ways of Semantic Knowledge Representation in Cognitive Architectures. ACT-R represent the semantic knowledge in units called chunks (Anderson, 1996), SOAR is based on the form of representation of ACT-R, where each chunk collects a number of pieces of environmental information in a single unit (Gobet et al., 2001). CLARION uses various representations according to the type of knowledge used: explicit objective knowledge is symbolic, while implicit procedural knowledge is sub-symbolic (Kotseruba, Gonzalez, & Tsotsos, 2016).

Recent researches on cognitive architectures show some representations of knowledge, for example, Vector LIDA (Snaider & Franklin, 2014) employs high-dimensional vectors. High-dimensional vector spaces have noteworthy properties that make them attractive for representation models (Snaider et al., 2014). These vectors have interesting properties that make them attractive for representations in cognitive models. LEABRA (O'Reilly, Hazy, & Herd, 2015) are based on a purely connectionist model that allows the representation of knowledge. This model represents a worthy explanation of connections for the functional structure and interaction of human attentional networks (H. Wang & Fan, 2007), the early versions of cognitive architectures did not have a powerful semantic memory system that facilitated the development of search processes but rather used relational databases displaying an interface to communication, which made these search processes slower and more complicated. In cognitive architectures the development of cognitive agents requires experts in the language of each architecture to establish the way of representing knowledge and can reason about the world based on a system of semantic memory. The creation of cognitive agents is a difficult task and there is no standardized way to develop them, for that reason, CARINA uses different cognitive approaches such as self-regulation, metacognition and metamemory.

Nowadays it is a difficult and time consuming task to use cognitive architectures for the development of cognitive agents because this type of assignment requires expertise in the own language of that architecture and a very particular knowledge about forms of semantic representation of the architecture. In addition, there is no a systematized way in cognitive sciences neither in computer sciences to design cognitive agents even more if these agents are to perform self-regulation, metacognition and metamemory. The focus of this paper is to present the appropriate Semantic Knowledge Representation of a cognitive model in CARINA. Currently, the metacognitive architecture CARINA executes cognitive

models at the object level using its working memory and it has not mechanisms that allows to store them in its semantic memory. For this reason, the main contribution of this paper is to introduce a new approach for Semantic Knowledge representation of any type of cognitive models which CARINA could be able to execute them using its semantic memory. The CARINA metacognitive architecture requires very particular specifications because it is focused on metacognition, processes of self-regulation and metamemory, which handle chunks and representations commonly used in cognitive architectures that do not take into account the structure of the self-model and the Meta knowledge aspects. This chapter shows the representation of semantic knowledge in the form of beliefs of a cognitive model in metacognitive architecture CARINA. The semantic knowledge in CARINA presents the states in which the system is found. In CARINA Beliefs are information of facts or knowledge that create the Semantic Memory in CARINA. Also, shows an example of the representation of the semantic knowledge of a cognitive model. This chapter is organized as follows: first, a description of the metacognitive architecture CARINA is presented. Then, the structure of a cognitive model in CARINA is formally described using denotational mathematics and the Formal Representation of Semantic Memory Units (SMU) in CARINA. Finally, the implementation and comparison with other cognitive model architectures in CARINA is described.

BACKGROUND

A cognitive model is a computational model of some internal information processing mechanisms of the brain used in the area of Artificial Intelligence for the purposes of comprehension and prediction (Levesque, 2004). Cognitive models are developed to study different aspects of cognition: attention and multitasking, judgment and choice in decision-making and skill acquisition in dynamic situations (Jerónimo, Caro & Gómez, 2018). In the literature has been found different studies about the use of Cognitive Architectures in the Cognitive Models design process as follows: Peebles & Cheng (2002) showed a cognitive model to explain the response time and accuracy of an experiment in which participants' eye movements were stored while various problems were solved with different types of graphs. This cognitive model was based on ACT-R and it allowed to describe a perceptual pattern of ocular fixation by perceptual chunks over time. Thus, fixation patterns were obtained reflecting the limitations of working memory during problem solving. Ritter, Anderson & Koedinger (2007) implemented a cognitive tutor for the mathematics teaching process through the incorporation of a cognitive model in the ACT-R architecture taking as a source of information the learning results, memory and problem solving of the students, with the objective of interpreting the behavior and developing better plans of classes, activities and experiences to learn mathematics.

MacDougall (2016) designs a multi-method approach to model expert cognition in naturalistic environments, based on expert knowledge and existing methodologies for the design of cognitive models. The authors used three methodologies as follows: i) *Macro-SGOMS cognitive model*: This methodology uses theoretical basis of Sociotechnical Goals, Operators, Methods, Selection rules to generate a cognitive model of a cognitive task; ii) *Communication Analysis - Coordination and the division of labor*: This type of methodology was used to enhance the individual SGOMS model to a team model by examining coordination, communication and division of labor in expert communities; iii) *Computational implementation*: In this stage, a implementation in a computer system of the primary points of interest was developed using Python ACT-R for modeling the SGOMS framework, showing team-based coordination units and an interruption mechanism that allowed computational agents to respond without collapsing. Fu & Pirolli (2007) describe a computational cognitive model that explains navigation on the World

Wide Web. The model evaluates the perceived relevance of a web link for a user by using sequential link propagation mechanisms. The model was tested in the ACT-R architecture evaluating the behavior of eight subjects while searching for information on the Internet. The model allowed to dynamically update the evaluation and to decide when to leave a page based on previous experiences providing a good interaction between user and the web.

Nuxoll & Laird (2004) show a Cognitive Model of Episodic Memory Integrated with the SOAR cognitive architecture where the functional steps of the episodic memory to recover the information of facts are described. For this reason general automatic mechanisms are created to remember and to accede to the past to improve the behavior of cognitive agents. Also Mendonca & Wallace (2007) develop a cognitive model implemented in the architecture SOAR for improvisation in emergency situations in which they explain how to act quickly in complex situations, eliminating extensive planning in emergency management and symbolically implementing a theory of cognition in emergency management, based on references to respond to non-planned situations. According to Benjamin, Lyons & Lonsdale (2006), current cognitive architectures for running cognitive models in robotics lack support for concurrence and active perception of environmental elements. For this reason the ADAPT project (Adaptive Dynamics and Active Perception for Thought) is created that is able to model the world as a network of concurrent schemes understanding perception as a problem-solving mechanism, adapting the semantics of natural language to control the actions exercised by the robot. ADAPT was implemented in the SOAR architecture to manage the complexity of its environment and learn new schemes from experience. Kaminka & Fridman (2006) designed a cognitive model of Crowd Behavior Based on the Festinger's Social Comparison Theory (SCT), which explains that people tend to compare their behaviors with others who are more like them. The cognitive model was created for the SOAR cognitive architecture presenting an algorithmic framework which evaluates its implementations in several crowds under a computerized system that when executed individually acts by many agents that behave like humans.

However, the models described above only limited to a numerical or computational description of the most relevant aspects in the resolution of a cognitive task. All these studies, although they use cognitive models in cognitive architectures, do not have a semantic representation that allows them to be stored permanently in the long-term memory of the agent or intelligent system that executes it, for this reason, the work presented here makes an approximation from the description of the fundamental processes that underlying the development of a cognitive computationally task through the formal representation of a cognitive model with the use of denotational mathematics, and additionally shows a semantic representation of a cognitive model described as a result of CARINA metacognitive architecture.

The Semantic Knowledge Representation is a modeling and presenting method of knowledge that uses tools for representing from notions, concepts and formal symbols to collections of propositions (Rajman & Chappelier, 2010). There are several ways to Represent Semantic Knowledge: rules, tables and frames (Tanaka, Aoyama, Sugiura, & Koseki, 1995), Paradigms, facets and Trees (Kwasnik, 1999), ontologies (Chen, Liao, & Prasad, 1998);(Garcia-Crespo, Ruiz-Mezcua, Lopez-Cuadrado, & Gonzalez-Carrasco, 2011), metadata (Matta, Corby, & Prasad, 1998), agents (O'Leary, 1998), Semantic Networks (Peters & Shrobe, 2003);(Miller, 1995), Chunks (Anderson, 1996);(Avery, Kelley, & Davani, 2006), Graphs (Arevalillo-herráez, Arnau, & Marco-giménez, 2013) and neural networks (Zhou, Zhou, He, & Wu, 2016) among others.

In the literature exists different studies that use Semantic Knowledge Representation in Artificial Intelligent Systems. (Peters & Shrobe, 2003) present a semantic network as a way to represent knowledge by constructing intelligent spaces to encapsulate rooms, users, groups, roles and other type of information, which is a fundamental design issue. They use the semantic network structures to capture meetings as

they occur, linking together the main meeting topics along with their contributors and attendees. Given the nature of semantic networks, the addition and change of information are localized and clear, and the generation of inference can be done easily. They are exploring different ways to enrich the interactions the user can provide. They are including a network query language and an activation system. That allows a user to specify a segment of the network that they are looking for, and to receive a notification every time that network element is created.

Other example of most common and widely used semantic knowledge modeling tool is WordNet. WordNet (Miller, 1995) is the resource used to identify shallow semantic features that can be attached to lexical units. (Scott & Matwin, 1998) describes experiments in Machine Learning for text classification using a new representation of text based on WordNet hypernyms that can lead to significant loss in error rates in certain types of text classification tasks. This method uses the lexical and semantic knowledge incorporated in WordNet to move from a bag of words based on hypernym density. The appropriate value for the height of the generalization value depends on the characteristics of each classification task. Another benefit of hypernyms density representation is that the rules are often simpler and more understandable than the rules induced using the word bag.

(Avery et al., 2006) developed the architecture of the Symbolic and Intelligent Robotic Intelligence Control System (SS-RICS). SS-RICS contains production systems, semantic networks, machine learning and sub-symbolic processing to perform the control of the mobile robot in real time. These authors showed the integration of cognitive psychological techniques and artificial intelligence (AI) techniques for allowing to the mobile robots to understand and interacting of a superior way with the dynamic world in which we live. The system was inspired by ACT-R and other computational cognitive architectures that are attempting to merge distributed and symbolic AI approaches. SS-RICS concentrates on using an approach with particular strengths at the right layer in a multi-layered approach. They are leveraging the interaction of different knowledge representation algorithms, more so than they are stressing the use of any particular knowledge representation algorithm.

(Zhou et al., 2016) presents a new perspective to learn the semantic representation of queries and responses in the use of a neural network architecture. The characteristics learned from the semantic level are finally incorporated into a learning framework to classify. The network determines the semantic relation for the query-response pairs. Then, the semantic level characteristics are added in a framework, as well as a set of statistics-based features. The importance of each characteristic in comparison with the rest is learned through a Ranking algorithm. The experiments they did in both the automatic evaluation and the manual evaluation showed a significant improvement in the recovery of questions from the community.

However, in the first versions of cognitive architectures they did not have a powerful semantic memory system that allowed searches, but used relational databases and created an interface to communicate, which made it slower and more complicated. And there was not a Semantic Representation of cognitive models that are executed in those cognitive architectures.

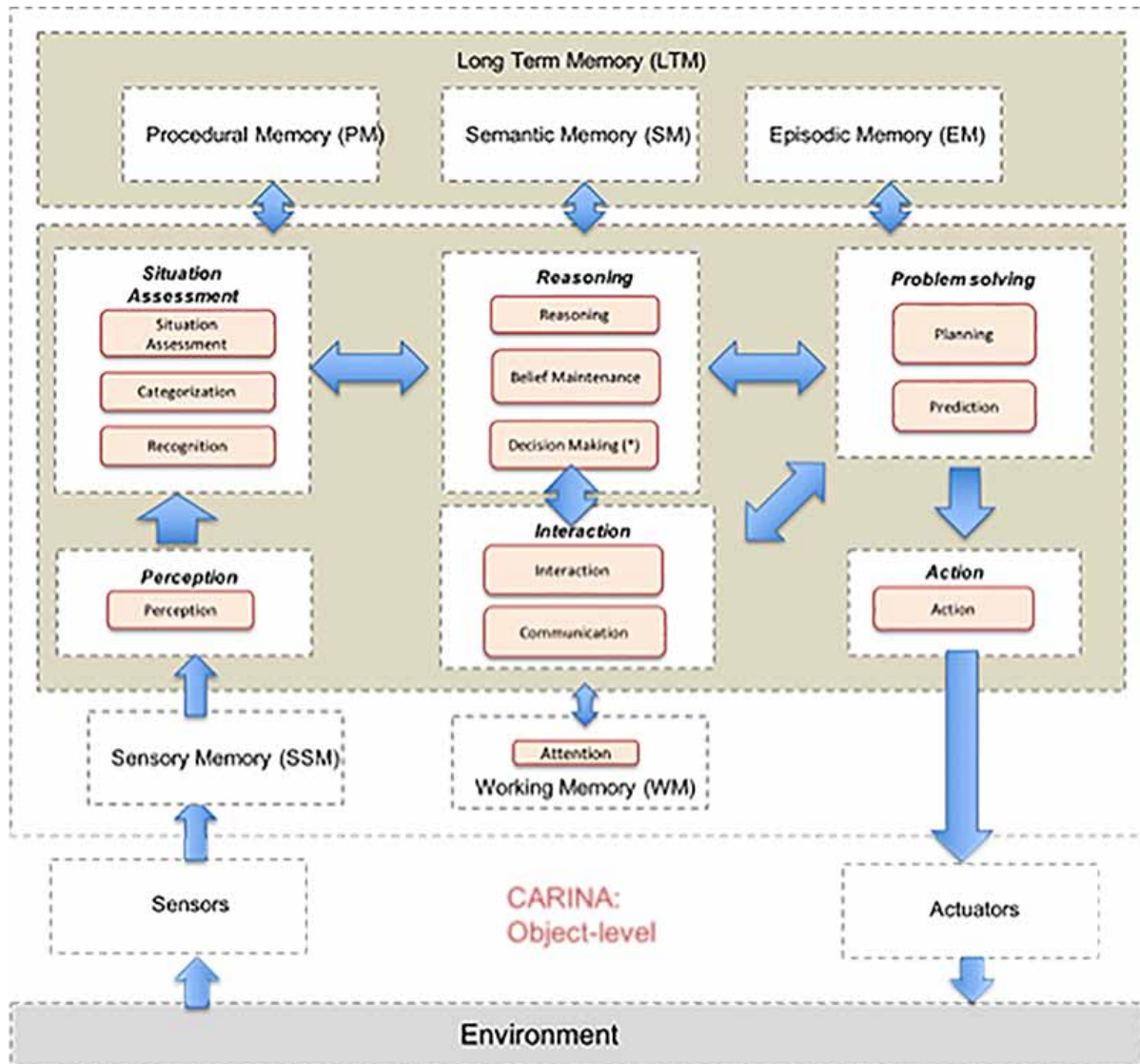
FOCUS OF THE ARTICLE

CARINA Metacognitive Architecture

CARINA is a metacognitive architecture for artificial intelligent agents. CARINA is derived from the MISM Metacognitive Metamodel (Caro et al., 2014). CARINA integrates self-regulation and metamemory with support for the metacognitive mechanisms of introspective monitoring and meta-level control; in

this sense CARINA assumes a functional approach to a mind philosophy, according to (Fodor, 1975), (Piccinini, 2010), (Scheutz, 2001). (Machamer, Darden, & Craver, 2000), the term “mechanism” includes entities and activities involving entities (i.e., including both static and dynamic aspects). The entities in CARINA are called “cognitive elements”. In accordance with MISM, CARINA is composed of three types of cognitive elements: structural elements, functional elements and basic elements. Structural elements are containers into which the functional and basic elements are embedded; the main structural element is the “Cognitive level”. The functional elements are tasks that enable reasoning and decision-making. Basic elements consist of the set of elements that participate and interact in reasoning and meta-reasoning processes. The main functional elements in CARINA are: reasoning task and meta-reasoning task. Reasoning Tasks are actions that enable the processing (transformation, reduction, elaboration, storage and retrieval) of information by applying knowledge and decision making in order to meet the

Figure 1. Structural view of CARINA (own source)



objectives of the system. A Metareasoning Task may be used to explain errors in some reasoning task or it may be used to select between “cognitive algorithms” to perform the reasoning (Raja & Cox, 2011).

CARINA is composed of two cognitive levels named object-level and meta-level. On the one hand, the object-level contains the model that an artificial intelligent agent has for reasoning about the world (i.e. agent’s environment) to solve problems (Caro et al., 2018). Object-level behavior consists of cognitive functions such as problem solving or memory retrieval (Cox, 2005). In CARINA, object-level has stages which are sets of cognitive functions (CF), (See Figure 1), Perception Stage (perception CF), Situation Assessment Stage (Situation Assessment CF, Categorization CF, and Recognition CF), Reasoning Stage (Reasoning CF, Belief Maintenance CF, and Decision Making CF), Problem Solving Stage (Planning CF, Prediction CF), Interaction Stage (Interaction CF, Communication CF) and Action Stage (Action CF). In this object-level the cognitive models designed by cognitive modeler are executed. In CARINA, the cognitive model must be loaded in its working memory each time that is wanted running it. This occurs due to that CARINA does not have a mechanism to store the cognitive models that have already been executed in it.

On the other hand, the meta-level contains a dynamic model of the object-level (Cox, 2005). The meta-level includes the components, knowledge and mechanisms necessary for a system to monitor and control its own learning and reasoning processes.

The reasoning process in CARINA takes as input beliefs, assumptions and expectations. It then reaches some mental conclusion (e.g. categorization). This is a central cognitive activity that enables an agent to augment its knowledge state. The problem solving process is closely related to reasoning in the sense that it is given a goal (in the world or within the agent) and the agent then chooses and orders its actions to achieve this goal. Monitoring a plan’s execution can also lead to revised estimates about the plan’s effectiveness, and, ultimately, to a decision to pursue some other course of action. In CARINA, decisions are often associated with the recognition of a situation or pattern, and combine the two mechanisms in a recognized- act cycle that underlies all cognitive behavior. Finally, acting involves the representation and storage of motor skills that enable the activities stored in Procedural Memory.

Formal Representation of Cognitive Model in CARINA

According to Jerónimo et al., (2018) a cognitive models are established to explore various areas of cognition; attention and multitasking, judgment and choice in decision making and skill building in dynamic conditions (Gonzalez, 2002). Cognitive architectures have been utilized to developed cognitive models of a range of intelligent systems (Forstmann & Wagenmakers, 2015). A cognitive model of a task built in a cognitive architecture is runnable and provide a string of conducts (Byrne, 2003).

Here officially a cognitive model in CARINA expending items of denotational mathematics suggested by (Wang, 2015) is described. A cognitive model consists of declarative knowledge, that indicate facts, also a procedural knowledge indicating reasoning rules. In addition modeling how to reason, as well necessary to model the control on when to reason about what (Muller & Heuvelink, 2008)

Let a CM cognitive model in CARINA, with:

$$CM \triangleq (P, G, S, MS, PK, SK)$$

Where:

p is the problem to solve utilizing CARINA.

$g \in G$, where g is a goal. A goal is an aim the system under consideration should obtain
 $s \in S$, where S is a Sensor. A sensor has the role of inspect the profiles of cognitive tasks to detect irregularity that may exemplify reasoning failures occasioned by the cognitive task (Caro et al., 2014).
 $m \in MS$, where MS is a mental state. Mental state is a Booleans variable that could be right or wrong. A state of mind is a state in which an agent is found.
 PK , represents the procedural knowledge that the system needs to perform the cognitive task, this is how process knowledge is formed in CARINA.

Let: $r \in R$, r is a production rule, with

$r \in (\text{condition}, \text{conclusion})$

With

$\text{condition} \triangleq (AS, C)$

With

AS is a sequence of variables, used to execute rules that fire the rules established in the cognitive task, in this sense a rule r is fires if there is a complete match with AS variables

$c \in C$ and $c \neq C$ and $c \neq AS$

c represent an additional limitation appended by cognitive developer.

A constraint (c) is a particular condition to match, in some conditions $c = \{ \}$

C are the conclusions. Conclusions are the actions triggers when the rule (c) are fires

In this context $a \in A$, with a is an action.

SK is a Semantic Knowledge. The Semantic Knowledge is the knowledge necessary to conduct the cognitive task. Semantic Knowledge is a group of beliefs archive in semantic memory for recovery over reasoning process.

Let: $f \in F$ with f is a Field. Where

$FD \subset F \wedge FB \subset F$

And F is a set of attributes of a SMU

A SMU is the Semantic Memory Unit, which semantic memory is composed by:

FD are primitive group of data type like as string, integer.

FB are collections that aim to a Beliefs of semantic memory.

And β are the Beliefs. The beliefs are clear epistemic atoms that depict declarative knowledge (Pezzulo & Calvi, 2004). According to (Rao & Georgeff, 1991) beliefs get and store information generated by the environment. Beliefs in CARINA architecture are the items of declarative knowledge, thus, information based on events that are stored. So beliefs represent the minimum unit that structures semantic memory in CARINA architecture (Caro et al., 2018).

The cognitive models in CARINA start with the specification of the problem to solve. In the sum of two numbers, the problem has three fields that CARINA must obtain. CARINA can get information of the problem from the sensors, estimating or recovering the information from the long-term memory.

```
1 "PROBLEMS": {
2 "TYPE": "ADDITION-SINGLE-COL-PRO",
3 "ADDENDUM_1": "EMPTY",
4 "ADDENDUM_2": "EMPTY",
5 "SUM": "EMPTY"
6 }
```

When CARINA reads a problem, then it creates a problem space. For each problem space, CARINA inserts a group of mental states that enable it to have updated information about the current state of the missing information to solve the current problem.

```
1 "MENTALSTATES": {
2 "PROBLEM_IS_INITIALIZED": FALSE,
3 "ADDENDUM_1_IS_READ": FALSE,
4 "ADDENDUM_2_IS_READ": FALSE,
5 "SUM_IS_CALCULATED": FALSE,
6 "SUM_FACT_IS_SAVED": FALSE,
7 "SUM_IS_DISPLAYED": FALSE,
8 "PROBLEM_IS_DONE": FALSE
9 }
```

CARINA can create goals using the mental states of the problem space. Every goal is directed to reach shifts in one or some mental states. A goal has posited both the current value of the current mental state and the expected value after goal method execution. The following two code fragments shown an example of the knowledge composition that save a goal in CARINA. For example, goal g701 points to modify the addendum_1_is_read mental state. When the goal is obtained the internal state changes from False to True. In CARINA, the group of goals that the system generates within the current problem space is named Motivational System.

```
1 "G701": {
2 "MENTAL_STATE": "ADDENDUM_1_IS_READ",
3 "CURRENT_VALUE": FALSE,
4 "TARGET_VALUE": TRUE
5 }
```

A group of production rules and an inventory of available cognitive functions represents the procedural memory of CARINA. In this code fragment a production rule is displayed.

```
1 "RULE702": {
2 "CONDITION": {
3 "ATTENTION_SYSTEM": {
```

```

4 "PROBLEM": "P2",
5 "MENTAL_STATE": "ADDENDUM_2_IS_READ",
6 "GOAL": "G702",
7 "SENSOR": "SENSOR1"
8 },
9 "CONSTRAINT": {}
10 },
11 "CONCLUSION": {
12 "ACTION1": {
13 "NAME": "READNUMBERFROMPERCEPTOR",
14 "COMPLETE": FALSE
15 },
16 "ACTION3":{
17 "NAME": "NEXTGOAL",
18 "ARG": {
19 "GOAL": "G703"
20 },
21 "COMPLETE": FALSE
22 }
23 }
24 }

```

Fragment 1 includes the name of the rule that is a unique identifier. The condition part of the rule encoded from fragment 2 to 8. The condition includes the variables of the attention system that are important for the rule to be triggered in the context of the current problem.

The conclusion part of the rule covers lines 11 to 23. This fragment shows the order in which the basic cognitive functions of CARINA must be implemented. Fragment 12 and 16 show the calls to two functions `readNumberFromPerceptor` and `nextGoal`.

A cognitive model also includes semantic knowledge in the form of Beliefs. In the cognitive model, the Beliefs essential for CARINA to resolve problems in a knowledge dominance are specified. In the following extract of code the basic structure of a belief that represents the number 1 in the problem of the sum of two numbers is described.

```

1 "ONE": {
2 "TYPESMU ": "NUMBER",
3 "HAS": {
4 "TYPE": "INTEGER",
5 "VALUE": "1",
6 }
7 }

```

Comparison with Other Cognitive Architectures

This part explains the similarities and differences among the implementation of cognitive models in CARINA and ACT-R. ACT-R was used for analogy due to its disclosure and extensive number of published cognitive models. For the comparison, the model explained in paper (Lebiere, 1999) was taken.

Similarities

In ACT-R, declarative knowledge is described in the form of “chunks”, which are singly a set of key-value pairs. In CARINA, beliefs are utilized in the similarity of chunks. The difference is the internal structure of each one, since the beliefs of CARINA have two sections with two relationships: IS-A and HAS. The procedural knowledge in CARINA and ACT-R is represented in the form of production rules, although they do not have the same structure, they execute comparatively equal.

Differences

The definition of a problem in CARINA has a particular specialized structure, while in ACT-R the problem is described in the condition of one or more production rules, as shown in the example below.

```
1 (P INITIALIZE-ADDITION
2 =GOAL>
3 ISA ADD
4 ARG1 =NUM1
5 ARG2 =NUM2
6 SUM NIL
7 ==>
```

CARINA in addition to having production rules can have specialized algorithms that in a more granular form are triggered by the production rules.

Formal Representation of Semantic Memory Units (SMU) in CARINA

Cognitive agents must have the ability to represent knowledge, learn and reason which remain as research challenges in the area of agent world (Nwana & Ndumu, 1999). All the knowledge received by CARINA is stored in the semantic memory in the form of beliefs. The nodes found in the beliefs network refer to the semantic memory units (SMU), and links refer to the various relations between these units (Shi, Zhou, & Wang, 1997).

$\beta = \langle \text{ISA}, \text{HAS} \rangle$

Where $\beta \in B$ and β is a set of SMU, with:

ISA is a set of fields, where $f \in \text{ISA} \wedge f \in B \wedge \text{ISA} \neq \{\}$

HAS is a set of fields, where $f \in \text{HAS} \wedge f \in B \wedge f$ is a basic data type.

In formulae (4) β is a SMU that is the minimum unit of information contained in the semantic memory of a CARINA-based cognitive agent. According to (Shi et al., 1997) Semantic units are those components that are extracted from knowledge after it is “computed”. The set of fields C represent relationships of a hierarchical type between the SMU. A SMU can have one or several ISA fields; therefore it is necessary to know the context of each SMU for the disambiguation of the semantic relations within the declarative memory. The set of HAS fields can represent aggregation relationships (i.e. is-part-of links) or descriptive field relationships (qualities and characteristics of a concept).

Beliefs are explicit epistemic atoms that represent declarative knowledge. In CARINA, the basic units of representation of knowledge are beliefs; the beliefs are elements of declarative knowledge about facts, concepts or notions from the world and from same cognitive agent (Apt, Blair, & Walker, 1988). Beliefs constitutes the minimum unit of semantic memory.

According to the above, each Belief is structured into two essential parts: i) ISA: that specifies the category which belongs to this belief (and that points to another belief that is stored in the semantic memory); ii) HAS: which specifies the other fields that characterize this belief in particular. In the following fragment of code the basic structure of a belief that represents the number 1 in the problem of the sum of two is shown.

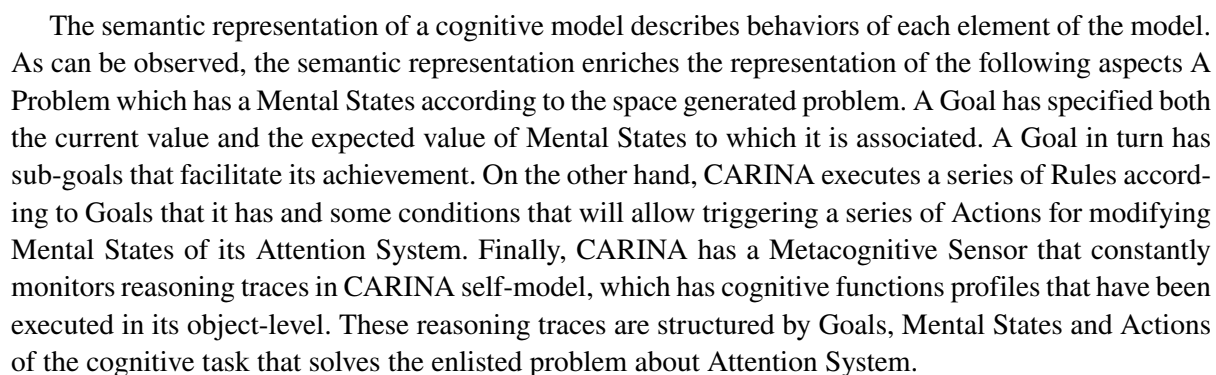
```

1 "one":{
2 "isa":{
3 "number"
4 }
5 "has": {
6 "type": "integer",
7 "value": "1",
8 }
9 }
```

Semantic Knowledge Representation of a Cognitive Model in CARINA

The Metacognitive Architecture CARINA executes cognitive models in the object level using its working memory. This cognitive models must be loaded each time that running them is wanted. CARINA has not mechanisms that allow store them it in its semantic memory. For this reason, it is necessary to semantically represent a cognitive model because after its execution, it can be stored in the semantic memory of CARINA and thus CARINA can execute it later on without direct supervision of cognitive designer. Thereby, the semantic representation of a cognitive model executed in the Metacognitive architecture CARINA is presented. This representation is made through of beliefs in Semantic Memory of CARINA, allowing thus a greater facility in designing the future learning processes of this Architecture. Taking into account theoretical assumptions of (Brachman & Schmolze, 1988), the semantic representation of a cognitive model in CARINA is useful due to fact that when it is represented symbolically, can be manipulated in an unsupervised way by the semantic memory of CARINA. Thus providing CARINA with the necessary foundations for the development of learning mechanisms that allow it to retrieve from its semantic memory each and every one of the cognitive models that have already been executed by it.

For this research we have represented these cognitive model elements (Problem, Goal, Sensor, Mental State, Procedural Knowledge and Semantic Knowledge) through of a Beliefs Map (See Figure 2).



The problem discussed in this chapter is the semantic representation of cognitive models executed in CARINA that allows store them in its long-term memory. For this reason is recommended develop learning mechanisms that facilitate the reusability and modularity of the constitutive elements of cognitive models. CARINA must be able to create its own cognitive models in an automatic way and it must have to a forgetting function that allow to liberate memory with repeated elements among cognitive models

FUTURE RESEARCH DIRECTIONS

In this work it is necessary to continue working on the representation of knowledge through the meta-cognitive architecture CARINA, for this reason:

- 1- CARINA must have a learning mechanism that allows it to compare the stored cognitive models, in addition must be able to save the necessary information from a cognitive model, such as: problem, goals, mental states, beliefs, actions and reasoning rules to solve and execute new cognitive tasks of the environment based on its previous learning.
- 2- Based on the stored cognitive models, CARINA must be able to create its own cognitive models to solve cognitive tasks in any environment.
- 3- CARINA must have a forgetting function for cognitive models that allows it to delete unnecessary information and optimize the algorithms that are executed.

This technology can be used for the development of information systems that automates tasks executed by human agents in organizational knowledge management processes using production rules. This computational systems that use cognitive models to simulate specific tasks in an organization, will allow in the future the development of intelligent agents that facilitate the execution of organizational processes in a most efficient and effective way in any type of environment.

CONCLUSION

In this chapter was presented a method for Represent Semantically a CARINA based cognitive model, this technique was called Beliefs. CARINA is a metacognitive architecture for artificial intelligence that was derived from the MISM Metacognitive Metamodel. The composition of a cognitive model was shown using denotational mathematics. A CARINA-based cognitive model includes the specification of a problem (cognitive task), a series of mental states that shows an initial state and a desired state, a series of goals, procedural knowledge and semantic knowledge. The cognitive task of adding two numbers, was used as an example, in which mental states, goals, rules of production and beliefs are described in detail. All the information received by CARINA is stored in the semantic memory in the form of beliefs. Therefore, all kinds of cognitive models represented through beliefs will be ready to be stored in the semantic memory of the metacognitive architecture CARINA. Carina has a representation in the memory of different types of structures of the cognitive model, in the form of beliefs. In this way CARINA in a near future, can autonomously generate a new complete cognitive model.

REFERENCES

- Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *The American Psychologist*, 51(4), 355–365. doi:10.1037/0003-066X.51.4.355
- Apt, K. R., Blair, H. A., & Walker, A. (1988). Towards a theory of declarative knowledge. In *Foundations of deductive databases and logic programming* (pp. 89–148). Elsevier. doi:10.1016/B978-0-934613-40-8.50006-3

- Arevalillo-herráez, M., Arnau, D., & Marco-giménez, L. (2013). *Systems Domain-specific knowledge representation and inference engine for an intelligent tutoring system*. doi:10.1016/j.knosys.2013.04.017
- Avery, E., Kelley, T. D., & Davani, D. (2006). Using cognitive architectures to improve robot control: Integrating production systems, semantic networks, and sub-symbolic processing. *15th Annual Conference on Behavioral Representation in Modeling and Simulation (BRIMS)*.
- Benjamin, D. P., Lyons, D., & Lonsdale, D. (2006, October). Embodying a cognitive model in a mobile robot. *Intelligent Robots and Computer Vision XXIV Algorithms, Techniques, and Active Vision*, 6384, 638407.
- Brachman, R. J., & Schmolze, J. G. (1988). An overview of the KL-ONE knowledge representation system. In *Readings in Artificial Intelligence and Databases* (pp. 207-230). Academic Press.
- Byrne, C. (2003). A unified treatment of some iterative algorithms in signal processing and image reconstruction. *Inverse Problems*, 20(1), 103–120. doi:10.1088/0266-5611/20/1/006
- Caro, M. F., Josyula, D. P., Gómez, A. A., & Kennedy, C. M. (2018, July). Introduction to the CARINA metacognitive architecture. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 530-540). IEEE. 10.1109/ICCI-CC.2018.8482051
- Caro, M. F., Josyula, D. P., Cox, M. T., & Jiménez, J. A. (2014). Design and validation of a metamodel for metacognition support in artificial intelligent systems. *Biologically Inspired Cognitive Architectures*, 9, 82–104. doi:10.1016/j.bica.2014.07.002
- Chen, Y.-M., Liao, C.-C., & Prasad, B. (1998). A systematic approach of virtual enterprising through knowledge management techniques. *Concurrent Engineering*, 6(3), 225–244. doi:10.1177/1063293X9800600305
- Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial Intelligence*, 189(2), 104–141. doi:10.1016/j.artint.2005.10.009
- Cox, M. T., Oates, T., & Perlis, D. (2011). Toward an Integrated Metacognitive Architecture. *AAAI Fall Symposium: Advances in Cognitive Systems*.
- Cox, M. T., & Raja, A. (2011). *Metareasoning: Thinking about thinking*. MIT Press. doi:10.7551/mitpress/9780262014809.001.0001
- Fodor, J. A. (1975). *The language of thought*. Thomas Y. Crowell.
- Forstmann, B. U., & Wagenmakers, E.-J. (2015). Model-based cognitive neuroscience: A conceptual introduction. In *An introduction to model-based cognitive neuroscience* (pp. 139–156). Springer.
- Fu, W. T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction*, 22(4), 355–412.
- Garcia-Crespo, A., Ruiz-Mezcua, B., Lopez-Cuadrado, J. L., & Gonzalez-Carrasco, I. (2011). Semantic model for knowledge representation in e-business. *Knowledge-Based Systems*, 24(2), 282–296. doi:10.1016/j.knosys.2010.09.006
- Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), 236–243. doi:10.1016/S1364-6613(00)01662-4 PMID:11390294

- Gonzalez, C. (2002). *The role of cognitive modeling in enhancing dynamic decisions*. Academic Press.
- Jerónimo, A. J., Caro, M. F., & Gómez, A. A. (2018). Formal Specification of cognitive models in CARINA. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 614–619). 10.1109/ICCI-CC.2018.8482062
- Kaminka, G. A., & Fridman, N. (2006, July). A cognitive model of crowd behavior based on social comparison theory. *Proceedings of the AAI-2006 workshop on cognitive modeling*.
- Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). *A review of 40 years of cognitive architecture research: focus on perception, attention, learning and applications*. CoRR Abs/1610.08602
- Kwasnik, B. H. (1999). *The role of classification in knowledge representation and discovery*. Academic Press.
- Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive Architectures: Research Issues and Challenges. *Cognitive Systems Research*, 10(2), 141–160. doi:10.1016/j.cogsys.2006.07.004
- Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft*, 8(1), 5–19. doi:10.1007/001970050071
- Levesque, H. J. (2004). *Knowledge Representation and Reasoning Acknowledgments*. Academic Press.
- MacDougall, W. K. (2016). *The Importance of Being Expert: A Multi-Method Approach to Modeling Expert Cognition in Naturalistic Environments* (Doctoral dissertation). Carleton University.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking About Mechanisms. *Philosophy of Science*, 67(1), 1–25. doi:10.1086/392759
- Matta, N., Corby, O., & Prasad, B. (1998). A generic library of knowledge components to manage conflicts in CE tasks. *Concurrent Engineering*, 6(4), 274–287. doi:10.1177/1063293X9800600401
- Mendonca, D. J., & Al Wallace, W. (2007). A cognitive model of improvisation in emergency management. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans*, 37(4), 547–561. doi:10.1109/TSMCA.2007.897581
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41. doi:10.1145/219717.219748
- Muller, T. J., Heuvelink, A., & Both, F. (2008, May). Implementing a cognitive model in soar and act-r: A comparison. *Proceedings of Sixth International Workshop "From Agent Theory to Agent Implementation"*.
- Nuxoll, A., & Laird, J. E. (2004, July). A cognitive model of episodic memory integrated with a general cognitive architecture. In ICCM (pp. 220–225). Academic Press.
- Nwana, H. S., & Ndumu, D. T. (1999). Perspective on software agents research. *The Knowledge Engineering Review*, 14(2), 125–142. doi:10.1017/S0269888999142012
- O’Leary, D. E. (1998). Using AI in knowledge management: Knowledge bases and ontologies. *IEEE Intelligent Systems & their Applications*, 13(3), 34–39. doi:10.1109/5254.683180

- Peebles, D., & Cheng, P. C. H. (2002). Extending task analytic models of graph-based reasoning: A cognitive model of problem solving with Cartesian graphs in ACT-R/PM. *Cognitive Systems Research*, 3(1), 77–86. doi:10.1016/S1389-0417(01)00046-8
- Peters, S., & Shrobe, H. E. (2003). Using semantic networks for knowledge representation in an intelligent environment. In *Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on* (pp. 323–329). 10.1109/PERCOM.2003.1192756
- Pezzulo, G., & Calvi, G. (2004). A pandemonium can have goals. *Proceedings of the Sixth International Conference on Cognitive Modeling*, 237–242. Retrieved from <https://pdfs.semanticscholar.org/b9da/007176b8be71370a0dfdec80dcdac47701e4.pdf>
- Piccinini, G. (2010). The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research*, 81(2), 269–311. doi:10.1111/j.1933-1592.2010.00356.x
- Raja, A., & Cox, M. T. (2011). *Metareasoning: Thinking About Thinking*. MIT Press.
- Rajman, M., & Chappelier, J.-C. (2010). *Constructing Semantic Knowledge Model based on Children Dictionary Masood Ghasemzadeh*. Academic Press.
- Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. *KR*, 91, 473–484.
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249–255. doi:10.3758/BF03194060 PMID:17694909
- Scheutz, M. (2001). Computational versus causal complexity. *Minds and Machines*, 11(4), 543–566. doi:10.1023/A:1011855915651
- Scott, S., & Matwin, S. (1998). Text classification using WordNet hypernyms. *Usage of WordNet in Natural Language Processing Systems*.
- Shi, Z., Zhou, H., & Wang, J. (1997). Applying case-based reasoning to engine oil design. *Artificial Intelligence in Engineering*, 11(2), 167–172. doi:10.1016/S0954-1810(96)00029-5
- Snaider, J., & Franklin, S. (2014). Vector lida. *Procedia Computer Science*, 41, 188–203. doi:10.1016/j.procs.2014.11.103
- Sun, R., & Naveh, I. (2004). Simulating organizational decision-making using a cognitively realistic agent model. *Journal of Artificial Societies and Social Simulation*, 7(3).
- Tanaka, M., Aoyama, N., Sugiura, A., & Koseki, Y. (1995). Integration of multiple knowledge representation for classification problems. *Artificial Intelligence in Engineering*, 9(4), 243–251. doi:10.1016/0954-1810(95)00006-5
- Van Lamsweerde, A. (2001). Goal-oriented requirements engineering: A guided tour. In *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on* (pp. 249–262). IEEE.
- Wang, Y. (2015). Concept Algebra: A Denotational Mathematics for Formal Knowledge Representation and Cognitive Robot Learning. *Journal of Advanced Mathematics and Applications*, 4(1), 61–86. doi:10.1166/jama.2015.1074

Zacarias, M., Magalhaes, R., Caetano, A., Pinto, H. S., & Tribolet, J. (2008). Towards organizational self-awareness: An initial architecture and ontology. In *Handbook of ontologies for business interaction* (pp. 101–121). IGI Global. doi:10.4018/978-1-59904-660-0.ch007

Zhou, G., Zhou, Y., He, T., & Wu, W. (2016). Learning semantic representation with neural networks for community question answering retrieval. *Knowledge-Based Systems*, 93, 75–83. doi:10.1016/j.knsys.2015.11.002

ADDITIONAL READING

Caro, M. F., Josvula, D. P., Gómez, A. A., & Kennedy, C. M. (2018, July). Introduction to the CARINA metacognitive architecture. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 530-540). IEEE. 10.1109/ICCI-CC.2018.8482051

Caro, M. F., Josyula, D. P., Cox, M. T., & Jiménez, J. A. (2014). Design and validation of a metamodel for metacognition support in artificial intelligent systems. *Biologically Inspired Cognitive Architectures*, 9, 82–104. doi:10.1016/j.bica.2014.07.002

Caro, M. F., Josyula, D. P., & Jiménez, J. A. (2014). A Formal Model for Metacognitive Reasoning in Intelligent Systems. *International Journal of Cognitive Informatics and Natural Intelligence*, 8(3), 70–86. doi:10.4018/IJCINI.2014070105

Gómez, A. A., & Caro, M. F. (2018, July). Meta-Modeling Process of Pedagogical Strategies in Intelligent Tutoring Systems. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 485-494). IEEE. 10.1109/ICCI-CC.2018.8482046

Jerónimo, A. J., Caro, M. F., & Gómez, A. A. (2018). Formal Specification of cognitive models in CARINA. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 614–619). 10.1109/ICCI-CC.2018.8482062

Olier, A. J., Gómez, A. A., & Caro, M. F. (2018, July). Cognitive Modeling Process in Metacognitive Architecture CARINA. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 579-585). IEEE. 10.1109/ICCI-CC.2018.8482090

Olier, A. J. O., Salgado, A. A. G., & Piñeres, M. F. C. (2017). Design and Implementation of a Teaching Tool for Introduction to object-oriented programming. *IEEE Latin America Transactions*, 15(1), 97–102. doi:10.1109/TLA.2017.7827913

KEY TERMS AND DEFINITIONS

Beliefs: Beliefs are a form of represent the semantic knowledge, which are elements about facts, concepts, or objects.

Cognitive Architecture: A cognitive architecture is a general framework for specifying computational behavioral models of human cognitive performance.

Cognitive Model: A cognitive model is a computational model of internal information processing mechanisms of the brain for the purposes of comprehension and prediction.

Denotational Mathematics: Is an abstract structure that uses exclusive semantics in almost all cognitive actions like those of thinking, knowledge representation, learning, and reasoning.

Metacognitive Architecture CARINA: CARINA is a metacognitive architecture for artificial intelligent agents. CARINA is derived from the MISM

Semantic Knowledge Representation: Semantic knowledge representation is a field of study which concentrates on using formal symbols to a collection of propositions, objects, object properties and relations among objects.

Semantic Memory: Semantic memory incorporates all achieved knowledge about the world, not contextualized in time and space.

Error Types in Natural Language Processing in Inflectional Languages

1

Gregor Donaj*University of Maribor, Slovenia***Mirjam Sepesy Maučec***University of Maribor, Slovenia*

INTRODUCTION

Natural Language Processing is the area of research that explores how humans can interact with computers in natural language. It is a subfield of Artificial Intelligence. The development of NLP applications is a challenging task, as natural languages are ambiguous, and not structured as strictly as programming languages, which were developed for the purpose of human-computer interaction. Today, we are faced with large amounts of data, available on the internet. The development of Natural Language Processing applications provides invaluable insight into very diverse data sources, which are not manageable by humans, but by the use of technology.

There are many different tasks in Natural Language Processing. To cover all of them within this article would be impossible. This article focuses only on two prominent Natural Language Processing applications: Automatic speech recognition and machine translation. The task of automatic speech recognition is for a machine to recognise the spoken words from an audio signal. This can be done on pre-recorded speech or live from a microphone. Essentially, it is the conversion from speech to text. The application area of automatic speech recognition is very broad. For example, it can be used for subtitling TV programmes for deaf people. Machine translation is the task of translating some text from one natural language to another. It can be applied to catch the meaning of a text written in a language that the reader does not understand.

The effectiveness of Natural Language Processing methods varies greatly, depending on the language under consideration. Highly inflected languages belong to a group of the most difficult languages to process. This article deals with one representative of this group, Slovene. It is a morphologically rich language with complex grammar, which harms the performance of Natural Language Processing tasks. The motivation of this article is to analyse errors that result from the characteristics of Slovene, and are not so frequent in Natural Language Processing of other languages.

BACKGROUND

Natural Language Processing

Large amounts of data are being created online every day, and many data are in the form of text in a natural language. The idea is to process and examine the data and to uncover new knowledge. Computers know how to process structured data, but data in natural language is unstructured. It requires specialised

DOI: 10.4018/978-1-7998-3479-3.ch006

approaches to process it. The research field that copes with this phenomenon is called Natural Language Processing.

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence. It focuses on enabling computers to understand and process natural languages as humans do. Although NLP research has a long history, many problems are still unsolved. Computers are far behind human abilities. We still do not know precisely how humans process language. Despite that, today, computers offer many, quite useful applications that are based on natural language. The machine learning evolution spurred remarkable technology breakthroughs. NLP evolved from a time-consuming process where rules were handwritten by humans, to unsupervised learning, where computers learn from data by themselves.

There are many interesting tasks which are based on NLP:

- Document summarization: Automatically generating synopses of large bodies of text.
- Automatic speech recognition: Transforming voice into written text.
- Speech synthesis: Transforming text into voice.
- Machine translation: Automatic translation of a text (or speech) from one language to another.
- Sentiment analysis: Identifying the emotions and subjective opinions within large amounts of text.
- Semantic analysis: Interpreting human sentences logically.
- Natural language understanding: Transforming the meaning of a text into a structured semantic form.
- Natural language generation: Generating text from structured data in a readable format with meaningful phrases and sentences.
- Question Answering: Generating answers to questions in the form of a sentence. It is based on natural language understanding.
- Dialogue systems with virtual assistants: Using natural dialogue that mimics a live agent interaction.

Every day, new ideas for applying NLP arise. The goal of almost all NLP tasks is to take raw language input (in written or spoken form) and use linguistic knowledge and algorithms to deliver higher value to the user.

In the continuation of this article we will focus on two NLP tasks: Automatic Speech Recognition and Machine Translation.

Automatic Speech Recognition

The most natural, and also essential, form of human interaction is by speech. In the modern world, however, more and more of our interaction is not only with other people, but with information systems. In the pursuit of a natural interaction between humans and machines, Automatic Speech Recognition (ASR) systems have been developed for the better part of the 20th century in order to be complementary with traditional methods of human-computer interaction, e.g., keyboards and graphical user interfaces. Today, the advancement of speech recognition can already be observed in commercial applications, such as home assistants.

Other applications of ASR may be the transcription of spoken dialogue into text form for the hearing impaired and speech to speech translation, where speech recognition is performed first, then machine translation, and then speech synthesis, in order to enable spoken interaction between people who do not speak the same language.

The vocabulary of an ASR system determines the set of words that can be recognised. Given the target application, vocabularies can have a size of several words to a few hundreds of thousands of words.

Simple speech recognition systems are able to recognise only a limited number of words spoken in an isolated manner, i.e., with pauses between words, while more complex systems can recognise words from a much larger vocabulary that are spoken continuously as a sentence.

In the process of speech recognition, firstly mel-frequency cepstral coefficients, or perceptual linear coefficients, are used typically for the extraction of relevant acoustic information from the audio signal in the form of features. Hidden Markov models are then used for acoustic modelling of phonemes and, sometimes, in systems with a limited vocabulary, modelling of whole words. Additionally, phonemic dictionaries are necessary to recognise words from a set of recognised phonemes. In such systems, only isolated words or phrases can be recognised.

In systems where continuous speech, i.e., naturally spoken sentences, are to be recognised, some kind of language model is necessary. In more limited systems, this can be achieved using regular grammar. This kind of language model is built by determining manually rules that describe allowed word sequences, and is implemented in the speech recognition system by the developers. This is feasible in more limited applications, designed for a limited number of usage scenarios.

Statistical language models are used for speech recognition in a more general domain. Those are built using appropriate tools from large training corpora, i.e., large collections of written texts. Statistical language models determine the probability that a given sentence or sequence of words might occur in the given language. Language models are not perfect, as they may assign a very low probability to sentences which happen not to occur in the training corpus. Also, as large corpora consist of written text, language models perform rather poorly while trying to model spontaneous speech, where a speaker might repeat or omit a word or correct themselves.

There are different challenges in different application areas of ASR. The first challenge is an accurate recognition of the spoken words. The accuracy of speech recognition systems is dependent greatly on the usage domain, the vocabulary size, the spontaneity of speech and the presence of background sounds.

How good is speech recognition? By far the most well-established metric for the evaluation of speech recognition systems is the Word Error Rate (WER). It is defined as the ratio between the number of (word level) errors and the total number of words in the correct sentence or speech segment which was processed by the system. There are three main types of errors: The deletion of a word, where the output of the system is missing a word; the insertion of a word, where a not spoken word appears in the system output; and the substitution error, where the system misrecognises a word for another. The Word Error Rate is determined using the so-called edit distance or Levenshtein distance, and is usually presented as a percentage. The value range for the Word Error Rate is from 0% upwards, and can, in extreme cases, even exceed 100% if there are more errors than actual words spoken. This would be possible if enough wrong words were inserted. Alternatively, the performance of an ASR system can be expressed with the accuracy which is calculated simply as the difference between 100% and the Word Error Rate percentage. A perfect speech recognition system has a Word Error Rate of 0%, or an accuracy of 100%.

Machine Translation

Language diversity in the world triggered the idea of Machine Translation to ‘overcome language barriers. To mention an example, the EU has 24 official languages, and it is very likely that one is confronted with a text in a language that he/she does not understand. Machine Translation (MT) is the automatic translation of a text, written in natural language, by a computer programme, with no human involvement.

The goals of the first MT systems were mainly to bring the meaning of some text to the reader who does not understand the language in which the text is written. Although today the goals of MT are set much higher, the primary goal is not always reached. We still do not have the warranty that the meaning of the source text is preserved in Machine Translation.

Human translation, however, often has a more ambitious aim, to produce texts that satisfy the linguistic norms of the target culture and are adapted to the assumed knowledge of its readers. From this perspective, MT will never supersede human translation. The research has shown that MT can produce human-like translations only in very restricted domains.

The first MT systems were rule-based systems. They appeared in the early fifties. Translation rules were created by experts of linguistics. The process was costly, and the rules were often not general enough to be applicable to other languages. Even for a predefined language pair, the sets of rules were hard to maintain and update. The first results of MT were very promising.

The idea of introducing statistics into Machine Translation has been studied since the early 1990s. Statistical MT (SMT) learns statistical dependencies from parallel and monolingual corpora. No linguistic knowledge is needed in the process of MT system building. Generally, SMT systems are not tailored to any specific pair of languages. To build an SMT system for a given language pair, only parallel corpora for that pair of languages is needed. If available, a monolingual corpus of the target language text can improve the fluency of Machine Translation. Translations in most successful SMT systems are based on phrases, defined by a data-driven process. An SMT system consists of many sub-components, each responsible for a specific part of the translation process, for example: For phrase translation; for reordering; for making the output a fluent sequence of words, etc. For a long time, SMT was by far the most widely studied Machine Translation method.

Neural Machine Translation (NMT) is the newest method of MT. The first experiments have shown its ability to create translations of higher quality than SMT. Unlike the traditional phrase-based translation system, which consists of many sub-components that are used as independent models, Neural Machine Translation attempts to build and train a single, large neural network that reads a source sentence and outputs a translation. Recurrent neural networks organised into an encoder-decoder architecture are used frequently. This architecture allows the variable length of source and target sentences as is common in translation.

How good is the Machine Translation? The evaluation of MT is not needed only to judge the translations of one system, but also to compare different MT systems, and to track the progress during the MT system building. Various methods for the evaluation for Machine Translation have been employed. Most reliable is the human evaluation, but it is costly and time-consuming. The automatic evaluation is an alternative. It is based on the correspondence between the MT output and reference translation (gold standard). Many evaluation metrics were defined by the research community. BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and TER (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006) are the most widely used. BLEU (Bilingual Evaluation Understudy) is the most commonly used evaluation metric. It measures the overlap of unigrams (single words) and high order n-grams between MT output and reference translations.

MT usually contains errors. Some of them can be fixed very easily. Because of that, the combination of MT with human post-editing of errors has become an appealing solution (Moorkens & O'Brien, 2015). It has been shown that post-editing MT can be faster than producing translations from scratch manually. It depends on the quality of translations produced by an MT system.

Inflectional Languages

The defining characteristic of inflectional languages is the use of an inflectional morpheme in its words to denote several of its grammatical features. Those can be tense, case, person, number among others. Word inflections are also present in non-inflectional languages. For example, in English, the suffix ‘-s’ is used to express nouns in plural form and the suffix ‘-ed’ to express verbs in the past tense. However, the degree of inflection in English is much smaller than in inflectional languages.

Examples of highly inflectional languages are Slavic languages. For instance, in Slovene, the verb ‘biti’ (to be) has 39 different forms. Other Slavic languages are Czech, Slovak, Polish, Russian, Ukrainian, Croatian, and Serbian, among others.

The rich morphology of inflectional languages results in greater challenges for NLP application. One simple factor is the high number of word forms, resulting in the need for larger vocabularies. For speech recognition, this results in a high number of acoustically similar words, that could be misrecognised. Another factor is a freer word order, which is problematic for Machine Translation, as the same meaning with the same words can be expressed differently. In both applications, the large number of words in the vocabulary causes a need for even greater training corpora for building language models.

Table 1 and Figure 1 show examples of the properties of words and sentences in Slovene. In Table 1 we see the noun ‘besedilo’ (text) in all 6 cases and three numbers of Slovene. The word has eight possible endings (including no ending in the form ‘besedil’). Additionally, the verb ‘besediti’ (to text, to put into words) has the same root and similar word forms as the noun. The verb has 16 different forms, of which some are the same as for the noun. As some word forms are still the same, their exact part of speech and morphological properties can only be determined by looking their context (the entire sentence).

Table 1. The word ‘besedilo’ (text) in all 6 cases and 3 numbers of Slovene

Case	Singular	Dual	Plural
Nominative	besedilo	besedili	besedila
Genitive	besedila	besedil	besedil
Dative	besedilu	besediloma	besedilom
Accusative	besedilo	besedili	besedila
Locative	besedilu	besedilih	besedilih
Instrumental	besedilom	besediloma	besedili

Figure 1 shows the example of a sentence in 2 different forms with different words orders. In the first example, the verb ‘je napisal’ (wrote) is before the subject ‘besedilo’ (the text), while in example 2, the subject is before the verb. Both sentences are grammatically correct and are also stylistically acceptable, while in English the sentence ‘The text he wrote today’ is not.

Figure 1. An example of the free word order in Slovene with 2 sentences with the same meaning and their English translation

Example 1: Danes je napisal besedilo.	(word order: verb, subject)
Example 2: Danes je besedilo napisal.	(word order: subject, verb)
English: He wrote the text today.	

We can also see that the Slovene word ‘danes’ (today) is at the start of the sentence, while its translation to English is at the end of the sentence.

Also the verb ‘je napisal’ is in the 3rd person masculine form, meaning that the Slovene examples do not need a pronoun to make the meaning of the sentence clear, while the English translation needs the pronoun ‘he’.

CLASSIFICATION OF ERRORS IN NLP APPLICATIONS

Identification of errors is an important step in every language-processing task. If we detect errors and analyse them, we can, consequently, improve the core procedures of the language processing task, or add an additional post-processing step in which we correct errors. In the continuation of the chapter, we focus on error analysis in ASR and MT.

Errors in Speech Recognition

The basic error types in Automatic Speech Recognition are deletion, insertion, and substitution. The Word Error Rate (WER) is calculated using the equation:

$$WER = \frac{D + I + S}{N} \cdot 100\%.$$

where D is the number of deleted words, I is the number of inserted words, S is the number of substituted words, and N is the number of all words in the correct transcription.

Figure 2 shows examples of all three basic error types, where LAB denotes the reference transcription, i.e., the correct spoken sentence, and REC denotes the speech recognition output. In Figure 2.a the word ‘je’ is missing in the recognition output, thus representing a deletion error. Similarly, in Figure 2.b the word ‘sicer’ is an insertion error, and, in Figure 2.c, the mismatch of the words ‘droge’ and ‘drugi’ is a substitution error.

Figure 2. Examples of the three basic error types in speech recognition, (a) A deletion error, (b) An insertion error, and (c) A substitution error

- a)
- | |
|--|
| LAB: naša ekipa pa si je razmere ogledala v dobu in se na lastne oči |
| REC: naša ekipa pa si razmere ogledala v dobu in se na lastne oči |
- b)
- | |
|---------------------------------------|
| LAB: imela drugi prosti program |
| REC: imela sicer drugi prosti program |
- c)
- | |
|--|
| LAB: povedala nam je da droge ne uživa že osem mesecev in se pri tem |
| REC: povedala nam je da drugi ne uživa že osem mesecev in se pri tem |

The causes of recognition errors can be diverse. They can be caused by acoustic factors, the speaker, language factors, the limited vocabulary, or the spontaneous nature of speech.

Furui, Nakamura, Ichiba and Ichiba (2005) examined the effects of spontaneous speech on speech recognition results. Adda-Decker and Lamel (2005) showed that speech recognition for male speakers had a higher error rate than for female speakers. Various authors have also shown that error rate increases with fast or very slow speech, as well as for infrequent words (Fosler-Lussier & Morgan, 1999, Shinozaki & Shinozaki, 2001).

Goldwater, Jurafski and Manning (2010) describe several factors that contribute to error rates of certain words. These can be disfluency factors, such as fragments and repetitions, or prosodic characteristics, such as pitch, intensity, duration, and jitter. They also describe double confusable pairs, which consist of words that are not only acoustically similar, but also have similar probabilities in the language model.

The confusion of acoustically similar words is very present in inflective languages, where the inflection results in several word forms, which, typically, differ in the word ending. This increases the required number of words in a vocabulary that would cover the language sufficiently. It is estimated that, for inflectional languages, the vocabulary size should increase by a factor of 10 compared to an English vocabulary in order to cover the same amount of the language. While using smaller vocabularies, a certain percentage of error is unavoidable, as spoken words are not in the vocabulary. Increasing the vocabulary size, therefore, reduces the error rate. However, increasing the vocabulary size results in large language models and the need for a larger training corpus (Donaj & Kačič, 2017, Nouza, Nejedlova, Zdansky, & Kolorenc, 2004).

Rotovnik, Sepesy Maučec, and Kačič (2007), dealt with the problem of vocabulary size by using subword based models, where words were split into the stems and the endings. While a small decrease in Word Error Rate was observed, the resulting language models and search algorithm were more complicated.

Donaj and Kačič (2017) have presented a classification of errors in an inflective language based on lexical features. They have found that approximately 10% of errors are due to a falsely recognised ending, resulting in a substitution error, where both words in the substitution pair have the same lemma. Similar errors in the word ending can appear, where the lemma is also changed, e.g., a substitution of adjectives and adverbs with the same word root. They also found that short words, including pronouns, prepositions and conjunctions, numerals, and proper names, have higher error rates. The highest error rate was observed with proper names.

Errors in Machine Translation

Errors in MT differ from errors in ASR. In speech recognition, words are recognised one after another, and in this process, because of miss-recognition, some words are omitted, substituted by acoustically similar words, or inserted. In the translation process different phenomena can be found. In a correct translation words often change their place in an output sentence (see the word ‘danes/today’ in Figure 1); the same word could be translated to different words, depending on the context; some words should be omitted or inserted in some contexts. All these events cause errors when a machine performs the translation. Especially in statistical MT, where translations are learned from large parallel corpora by discovering statistical relationships between words, specific errors are hard to predict and fix.

BLEU, the most used metric for the evaluation of MT, is calculated by the following formula:

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \cdot \log p_n \right).$$

where p_n is n-gram precision, and w_n is a positive weight. All weights sum to 1. N is the maximum order of n-grams. It is commonly set to 4. BP denotes the brevity penalty factor. With this brevity penalty in place, a high-scoring candidate translation must now match the reference translations in length. The brevity penalty is 1.0 when the candidate's length is the same as reference translation's length. Scores, between 0 and 100 (perfect match), are then averaged over the whole set of sentences selected for evaluation, to reach an estimate of the translation's overall quality.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. METEOR has additional features, i.e., stemming, paraphrasing and synonym matching.

TER (Translation Edit Rate) is a derivate from the WER. It uses an additional edit step, namely shifts of word sequences. A shift moves a contiguous sequence of words within the evaluated translation to another location within the translation. In TER all edits have equal cost. Although much effort has been invested in the automatic evaluation of MT, it is still difficult to trace the improvements in the MT system by computing the values of evaluation metrics. It is also difficult to say which MT system is better than others, based solely on values given by evaluation metrics. Even if a metric correlates well with human evaluation in one study on one evaluation set, this correlation may not carry over to other evaluation sets. Manual inspection is always beneficial.

It is important to know that some errors could happen in any Machine Translation task regardless the languages in translation. On the other hand, there are errors that occur in one language and not in another. In the continuation of the chapter, we focus on translation errors that appear if we translate between English and the highly inflected Slovenian language.

The most frequents issues when translating to and from Slovene are (Popović & Arčan, 2015):

- Multi-word expressions: Multi-word expressions consist of a head noun and additional nouns and adjectives. Multi-word expressions in English are frequently translated to unintelligible sequences of words.
- Negation: In English, negative clauses and sentences commonly include the negative particle not or other negative words, like no, nothing, never, etc. On the contrary, Slovene conventionally employs double negatives, where a negation is added to the verb and to the noun. Because of that, negation structures are translated incorrectly.
- Word order: If phrase boundaries are not detected correctly, words in translation are misplaced.
- Prepositions: There are hardly any rules as to when to use which preposition in English. A 1:1 translation to/from Slovene is usually not possible. One preposition in Slovene might have several translations in English, depending on the situation. Therefore, prepositions are frequently mis-translated, omitted or added.

Some translation errors depend on translation direction. The most frequents issues when translating to English are (Popović & Arčan, 2015):

- Articles: In Slovene, articles are not used. Because of that, some articles in English translation are missing.

- **Pronoun:** In some cases in Slovene, pronouns could be omitted. Therefore, several pronouns are missing in the English translation. The Slovene pronoun 'svoj' can be used for all persons, and it is difficult to disambiguate for translation in English.
- **Tenses:** The tense system in English differs to a great extent from tense systems in Slovene. Frequently, the tense from the source language is preserved and causes mistranslation.

For translation from English to Slovene, the problem is that, in the source sentence, much information is missing to generate rich morphology in translation. The most frequent errors are:

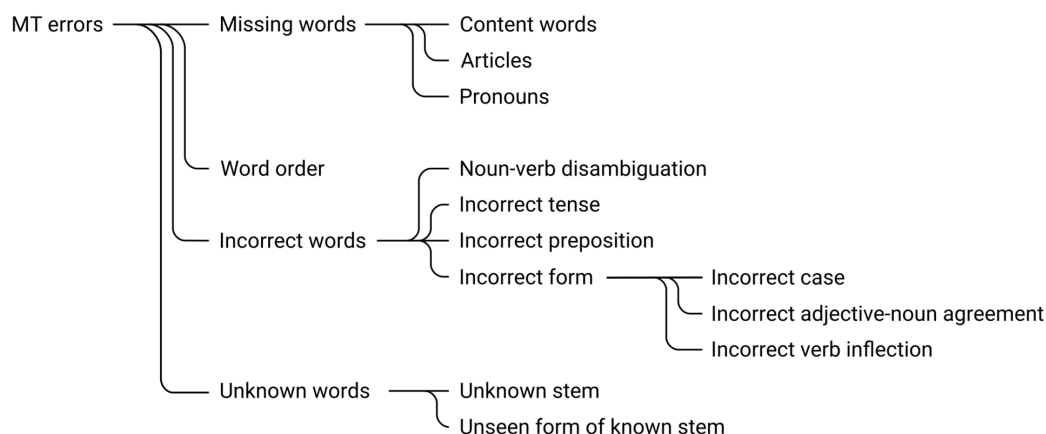
- **Case:** An incorrect case is used.
- **Adjective-noun agreement.** All Slovene adjectives must agree with nouns in number, gender, and case. Often incorrect cases appear in combination with incorrect number or gender agreements
- **Verb inflection:** Verb inflection does not correspond to the person, as in English there are only two verb inflections regarding the person. For a similar reason, sometimes, verb inflection does not correspond to the tense.
- **Noun-verb disambiguation:** There are a number of English words which can be used both as a noun as well as a verb. On the contrary, in Slovene, different words are used for nouns and for verbs. Sometimes it is difficult to disambiguate, and wrong words are chosen.

Error identification is not always a straightforward task. Sometimes errors are difficult to localise. Some errors are diffused across the sentence, or even a larger unit. Some examples of translation errors are collected in Table 2, while a more complete classification is given in Figure 3.

Table 2. Examples of MT errors

Error type	Sentence
Source sentence (English)	Being a rector does not give someone the freedom
Correct translation (Slovene)	Biti rektor nikomur ne daje svobode
Negation	Biti rektor ne daje <u>nekaj</u> svobode
Wrong case	Biti rektor nikomur ne daje svoboda
Wrong word order	Biti rektor nikomur ne <u>svobode daje</u>

Figure 3. Classification of error types in MT



SOLUTIONS AND RECOMMENDATIONS

In order to improve the performance of NLP applications, the types of errors have to be identified precisely enough to make an assumption for the reasons for the errors.

In Automatic Speech Recognition often only the word error rate is reported, with some distinctions in the test sets, e.g. acoustical quality or the gender of the speaker. Appropriate tools would be required to make a more precise analysis.

Such tools should be able to make a basic analysis of the types of errors (deletion, insertion, substitution). Then, a linguistic breakdown regarding the length of the words, their inflections, their part-of-speech (noun, adjective, verb ...), and other linguistic factors, might turn out to be useful. Additionally, speech rate, other prosodic factors, the level of background noise and other factors, could be included into the analysis by using the source audio signal or appropriate annotations of it.

When such a tool is available, several ASR systems might be tested and compared to make an estimate of which system performs better or worse with particular types of errors. Then, further research would be necessary to improve the performance of the ASR system with regard to particular types of errors for specific applications.

Otherwise, an error analysis tool would be usable, if it could provide an insight into the necessary steps and directions for error correction. This would enable post processing tools in which errors are further removed. It would also be useful to determine which ASR system results in recognition output with the least amount of work, if manual corrections would be applied.

The evaluation of Machine Translation output is an important and, at the same time, difficult task. A great deal of effort of the research community has been spent on finding automatic metrics which will enable ranking between different MT systems. Developed metrics correlate well with human judgments. It is useful, but to improve an MT system, additional information is needed in terms of errors that the MT system makes. This article presented an overview of the most common errors when translating a highly inflectional language. To judge a specific MT system, a systematic analysis is needed of errors in the system's output. Humans do it most commonly. The idea is to define automatic error analysis using external linguistic knowledge. It will provide more specific information of certain translation problems, on which the developer can focus and get ideas about possible improvements of the system. In (Popović & Ney, 2011) automatic error analysis has already been studied. It was based on position-independent WER in the first step to localise errors, and the application of linguistic rules in the second step to classify errors in different categories. A similar approach could be applied to highly inflected languages, as each language pair and each translation direction give rise to different translation errors.

Having the whole picture about translation errors of a specific MT system, automatic error correction methods should be investigated afterwards. It would probably be safer to apply error correction as a post-processing step. If we would try to model these aspects in translation models, changes in the model could trigger some other aspect of the translation in a negative way. Improvements in translation quality could also be expected if parallel corpora used for training MT systems had more training data.

MT systems are commonly developed for the general domain. As MT will always be error-prone, to limit the number of errors, MT can be developed as a domain-specific system. If we can detect the target domain of MT system usage, the system can be adopted for this specific domain. Consequently, the vocabulary is limited, and we avoid many mistranslated words.

Another method of improving the performance of NLP applications is their restriction or adaptation to specific domains. Domain knowledge increases the effectiveness of NLP methods. Domain specific methods reduce data sparsity and, consequently, improve the accuracy of the NLP task. The domain

of a target environment can be determined manually or detected automatically. Knowing the domain, domain-specific models are built. To train them, domain-specific data is needed. General-domain models are then adapted to a specific domain.

To build a domain-specific ASR system, a vocabulary and language model can be adapted to the target domain. The adaptation of vocabulary means that the vocabulary is extended with words specific to the target domain (for example with terminological terms). Also, words that do not occur in this domain might be excluded from the vocabulary. Using a fixed or adapted vocabulary, the language model can be adapted by training n-grams on domain-specific text. Those n-grams can define a component added to the general-domain language model.

The experiments show that general purpose translation systems usually perform poorly and, hence, it is important to develop translation systems for specific domains. The domain-specific MT is usually based on general-domain parallel corpora and domain-specific monolingual corpora. MT domain-adaptation can be data centric and model centric. Data centric methods focus on selecting training data from general-domain parallel corpora based a domain-specific language model. Model centric methods interpolate domain-specific and general-domain models in either a model level or instance level. Domain-adaptation has been studied widely for SMT, for the latest state-of-the-art NMT models it is a new and important future direction. Using domain specific dictionaries also improves MT significantly, as it solves many word sense disambiguation problems. How domain specific dictionaries could be used in NMT is also an open question. Using the unknown word replacement suffers from noisy attention.

FUTURE RESEARCH DIRECTIONS

Today, basic reserch tools for ASR and MT do not perform a thorough analysis of recognition or translation errors. Further research and development of appropriate tools would be an essential step, not only for the purpose of developing better applications, but also to understand the causes of errors and inner mechanisms of NLP applications.

Furthermore, the results of such an automatic analysis shall be used for adaptation or performacnce enhancement of NLP applications. This can also be done in conjunction with a manual error analysis. Then, those methods can also be adapted for inflectional languages by taking into account their morphological characteristics.

CONCLUSION

In this article, we presented an overview of errors that occur in NLP applications. Because natural language is a very complex phenomenon, it is not expected that in the future we could resolve all the errors occurring in NLP applications, but the number of errors could be reduced if they are analysed systematically. The article focuses on two NLP applications, that are used most widely in practice, ASR and MT. Challenges that we face in these two applications are different, although some types of errors occur in both of them. The difference between them also lies in the form in which the language is expressed. In ASR, it is a speech, and in MT it is a text. ASR copes with problems that arise in the acoustic realization of language, and MT with problems of textual realizations of natural languages. In the article, we presented the most common errors. Being aware of them is useful, because it gives ideas for further improvements of NLP methods. In the case of MT, error analysis is useful for the estimation of post-

editing effort. The question that needs a clear answer is if the effort to post-edit a sentence is expected to be lower than the effort to translate the sentence from scratch. The answer depends not only on MT errors but also on the translator's experience and skills. Nevertheless, it is an indicator of MT systems' usefulness. The article shows that there is still a lot of room for future research, which will make NLP applications even more attractive.

REFERENCES

- Adda-Decker, M. & Lamel, L. (2005). Do speech recognizers prefer female speakers?. In *Proceedings of interspeech 2005 - Eurospeech*. Lisbon, Portugal: ISCA.
- Denkowski, M., & Lavie, A. (2014). Meteor universal: language-specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, MD: Association for Computational Linguistics. 10.3115/v1/W14-3348
- Donaj, G., & Kačič, Z. (2014). Manual sorting of numerals in an inflective language for language modelling. *International Journal of Speech Technology*, 17(3), 281–289. doi:10.1007/10772-014-9231-y
- Donaj, G., & Kačič, Z. (2017). *Language modeling for automatic speech recognition of inflective languages: an applications oriented approach using lexical data*. Springer International Publishing. doi:10.1007/978-3-319-41607-6
- Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2-4), 137–158. doi:10.1016/S0167-6393(99)00035-7
- Furui, S., Nakamura, M., Ichiba, T., & Iwano, K. (2005). Why Is the Recognition of Spontaneous Speech so Hard? In V. Matoušek, P. Mautner, & T. Pavelka (Eds.), *Text, Speech and Dialogue (LNCS)* (Vol. 3658, pp. 9–22). Springer. doi:10.1007/11551874_3
- Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 55(3), 181–200. doi:10.1016/j.specom.2009.10.001
- Moorkens, J., & O'Brien, S. (2015). Post-editing evaluations: trade-offs between novice and professional participants. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT '15)*. Antalya, Turkey: EAMT.
- Nouza, J., Nejedlova, D., Zdansky, J., & Kolorenc, J. (2004). Very large vocabulary speech recognition system for automatic transcription of czech broadcast programs. In *Proceedings of interspeech 2004*. ISCA.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Philadelphia, PA: Association for Computational Linguistics.
- Popović, M., & Arčan, M. (2015). Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT '15)*. Antalya, Turkey: EAMT.
- Popović, M., & Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4), 657–688. doi:10.1162/COLI_a_00072

Rotovnik, T., Sepesy Maučec, M., & Kačič, Z. (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Communication*, 49(6), 437–452. doi:10.1016/j.specom.2007.02.010

Shinozaki, Z., & Furui, S. (2001). Error analysis using decision trees in spontaneous presentation speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01)*. Madonna di Campiglio, Italy: IEEE. 10.1109/ASRU.2001.1034621

Siegler, M. A., & Stern, R. M. (1995). On the effects of speech rate in large vocabulary speech recognition systems. In *1995 International Conference on Acoustics, Speech, and Signal Processing*. Detroit, MI: IEEE. 10.1109/ICASSP.1995.479672

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA '06)*. Oslo, Norway: AMTA.

ADDITIONAL READING

Costa, A., Ling, W., Luís, T., Correia, R., & Coheur, L. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, 29(2), 127–161. doi:10.1007/10590-015-9169-0

Donaj, G., & Kačič, Z. (2017). *Language modeling for automatic speech recognition of inflective languages: an applications oriented approach using lexical data*. Springer International Publishing. doi:10.1007/978-3-319-41607-6

Gales, M., & Young, S. (2008). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, 1(3), 195–304. doi:10.1561/20000000004

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 55(3), 181–200. doi:10.1016/j.specom.2009.10.001

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. doi:10.1017/CBO9780511815829

Olive, J., Christianson, C., & McCary, J. (Eds.). (2011). *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer-Verlag New York. doi:10.1007/978-1-4419-7713-7

Popović, M. (2018). Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, 32(3), 237–253. doi:10.1007/10590-018-9219-5

KEY TERMS AND DEFINITIONS

Automatic Speech Recognition: The automatic process of a machine to recognize spoken words from an audio signal or recording.

Inflectional Language: A language, where inflectional morphemes determine several morphological properties of a word.

Language Morphology: A subdiscipline of linguistics, which studies the formation of words in sentences.

Machine Translation: Translation of text or speech from the source language to the target language by a computer, with no human involvement.

Machine Translation Error: Error made by machine translation system. To some extent, machine translation errors differ from errors made by human translators.

Machine Translation Evaluation: Evaluation of translations, which resulted from Machine Translation. The evaluation can be manual or automatic. Automatic evaluation is done using an evaluation metric. The evaluation metric assigns a quality score to a translation by comparing it to the reference translation.

Natural Language Processing: A modern research area, dealing with the interaction between humans and computers with language.

Speech Recognition Error: The omission, insertion, or misrecognition of a word in the process of speech recognition.

Speech Recognition Vocabulary: The set of word a speech recognition system is able or designed to recognize.

Text-Based Image Retrieval Using Deep Learning

1

Udit Singhania <https://orcid.org/0000-0002-4400-2783>*Vellore Institute of Technology, India***B. K. Tripathy***Vellore Institute of Technology, India*

INTRODUCTION

This chapter is an advanced version of the previous chapter named ‘An Insight into Deep Learning Architectures’ in the book, ‘Encyclopedia of Information Science and Technology, Fourth Edition’. The earlier chapter was submitted by VIT University. It is already an established fact that information retrieval is of two types, image retrieval, and text retrieval. Image retrieval is further divided into a query and content-based approach. The main focus of the chapter will be the same as the earlier version, on Image Retrieval. In the previous edition of this book, the authors attempted to solve information retrieval problems using deep learning architectures. The authors looked on the RBMs, DBNs, and CNNs, how they can be used wisely to solve the issue of image retrieval from queries and text. This chapter will look at some of the advanced architectures (after RBMs and DBNs) used in the text-based approach.

What is information retrieval? Information retrieval means retrieving information from web or data present around us. Information retrieval is of two types image and text; information can be retrieved in either textual form or image form. Google Image search engines are an example of image retrieval. Retrieving documents from text is an example of text retrieval.

Image retrieval is done by deep learning architectures in action. Certainly, CNNs (convolutional neural nets) play an important role, they are the backbone of the image recognition tasks, with the help of NLP (natural language processing techniques) extracting the important features from the text-based query becomes an easy task. Feature learning is an important task while learning for images and text matching, this issue was resolved by using unsupervised approach DBN (Deep Belief Nets) and using new type of architecture was discovered named RBMs (Restricted Boltzmann Machines). After using simple RBMs, there has been a discovery of a new method named Multimodal Learning with Restricted Boltzmann Machines.

What is modality? In this world, information comes through various input channels. Some images have captions and tags, videos have visual and audio signals, sensory perception includes simultaneous inputs from visual, auditory, motor and haptic pathways. Each modality is characterized by separate statistical properties which makes it difficult to disregard the fact that they come from different input channels. Using the combinations of such modalities to jointly represent image and text helps in better models like for example, take an image of a car and words ‘A Car’, they are assigned a high probability to one conditioned on other. This approach is the base of Multi-modal RBMs.

In the supervised learning side, after CNNs came into existence, there have been dire need of an architecture which can remember past inputs and predict the future input and then came into existence,

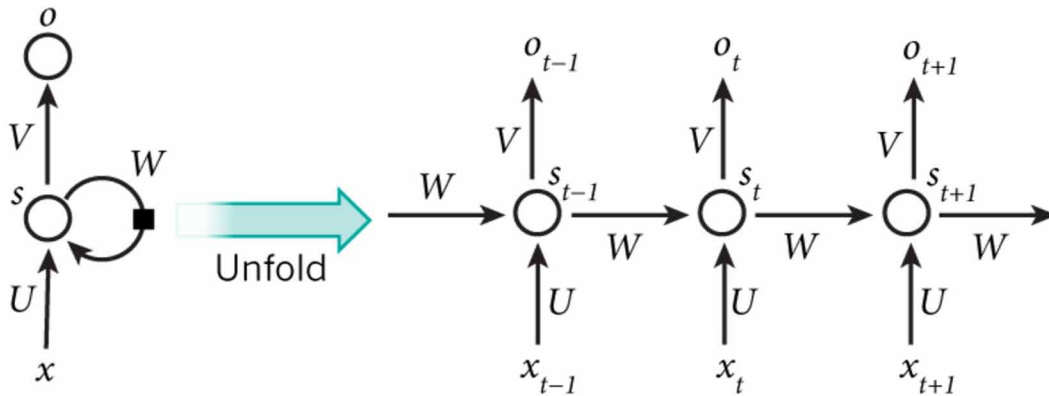
DOI: 10.4018/978-1-7998-3479-3.ch007

new architecture RNNs (Recurrent Neural Nets). RNNs play an important role in textual analysis, time series forecasting, sound and speech recognition.

BACKGROUND

Recurrent neural nets belong to the class of artificial neural networks (ANNs) designed specifically to recognize patterns in the sequences of data, such as text, handwriting, speech, or numerical time-series data which can be gathered from sensors, stock markets, and government agencies. These algorithms have a major focus on time and sequence, looking towards the temporal dimension. They are considered to be one of the most powerful and useful types of a neural network, alongside the attention mechanism and memory networks. RNNs are applicable even to images, which can be decomposed into a series of patches and treated as a sequence. Traditional neural nets could not use the sense of persistence, they failed to perceive the future input because of not able to remember the sequence in which the input was fed, these neural nets have basically two inputs, first they take the input of the previous activation function of the previous input and also consider the present input and they apply activation function of the resultant to get new activation function and predict the output. This can be understood by the figure drawn below.

Figure 1.



Here, RNN is being *unrolled* (or unfolded) into a full network. The formulas that govern the computation happening during a RNN area unit as follows:

- is the input at time step . x_0 is a one-hot vector with respect to 1st word of a sentence.
- is the hidden state at time step . It's the “memory” of the network. is calculated supported the previous hidden state and also the input at this step:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (1)$$

The function “f” usually is a non-linear such as tanh or ReLU (Rectified Linear Unit Function). S_{-1} , the first hidden state, is typically initialized to zero vector.

- O_t is the output at step t . In order to predict the next word in a sentence, result will be the vector of probabilities across the defined vocabulary $O_t = \text{softmax}(Vs_t)$.

RNNs started failing for longer time instances, for example,

“**Cat was** [playing with a ball of fur and] **it was** enjoying it very much.”

“**Cats were** [playing with a ball of fur and] **they were** enjoying it very much”

In the above example, there is a significant time gap between the “cat was” and “it was”. Hypothetically, RNNs are completely fit for taking care of such “long term dependencies.” A human could cautiously pick parameters for them to take care of toy issues of this structure. Tragically, in practice, RNNs don’t seem to be able to learn them. The main problem is of Vanishing Gradient. In Vanishing gradient problem, when dealing with activation functions like the sigmoid function,

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \quad (2)$$

Lies in [0,1] and when all the numbers between [0,1] are multiplied, the resultant gets smaller and smaller. So, the little estimations of function’s subsidiaries (happening in the mistake work) gets increased on different occasions as the user moves towards the beginning layers. Subsequently, the slope nearly vanishes as he moves towards the starting layers, and it winds up hard preparing these layers. This problem was resolved by a newer net called Long Short Term Memory nets which evolved with the idea of retaining the longer time gaps information so that they can be relayed successfully.

$$i = \sigma(x_t U^i + s_{t-1} W^i) \quad (3.1)$$

$$f = \sigma(x_t U^f + s_{t-1} W^f) \quad (3.2)$$

$$o = \sigma(x_t U^o + s_{t-1} W^o) \quad (3.3)$$

$$g = \tanh(x_t U^g + s_{t-1} W^g) \quad (3.4)$$

$$c_t = c_{t-1} \bullet f + g \bullet i \quad (3.5)$$

$$s_t = \tanh(c_t) \bullet o \quad (3.6)$$

LSTM is merely an another way to compute hidden state, just as equation (i) was used to compute in RNN. Now, looking at the above equations:

1. Here, i, f, o are the input, forget and output *gates*, respectively, they all have similar equations as in (i) but with different parameters. These equations are called *gates* because they use sigmoid function as an activation to squeeze the values of the vectors in the range of 0 to 1, and using the elementwise product with another vector they can be used to define the ratio of dominance of the other vector in the result. Following gates define the following:
 - A. Input gate defines the ratio of the newly computed state for the current input.
 - B. Forget gate defines the ratio of the previous state.
 - C. Finally, output gate defines the ratio of the internal state that is needed to expose to the external network.

All the gates have the same dimensions d_s , the size of the hidden state.

2. g is a “candidate” hidden state that is computed based on the current input and the previous hidden state.
3. c_t is the internal memory of the black box. It is the resultant of the previous memory c_{t-1} multiplied with forget gate, and the newly computed candidate state g , multiplied with input gate.
4. Using c_t , computing the output of hidden state s_t by multiplying the memory with the output gate. This gating mechanism was used to solve the problem of vanishing gradient.

MAIN FOCUS OF THE ARTICLE

Multimodal Deep Boltzmann Machine

Deep learning architectures like RBMs (Restricted Boltzmann Machines) have been used earlier in text based image retrieval, new type of architecture called Multimodal-DBM uses a twin type architecture to solve the problem. RBMs were originally developed for modelling binary vectors but **Gaussian-Bernoulli RBMs** are used in modelling pixel intensities and filter responses.

Energy function:

$$E(v, h; \theta) = \sum_{i=1}^D \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^D \sum_{j=1}^F \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_{j=1}^F a_j h_j \quad (4.1)$$

Distribution Function:

$$P(v; \theta) = \frac{1}{Z(\theta)} \sum_h \exp(-E(v, h; \theta)), Z(\theta) = \int_v \sum_h \exp(-E(v, h; \theta)) dv \quad (4.2)$$

And **softmax replication model** is used in word count vector representation. This model is a type of Restricted Boltzmann Machine in which the binary variables, have been replaced by multinomial one of a number of different states. Energy Function:

$$E(V, h) = -\sum_{i=1}^M \sum_{j=1}^F \sum_{k=1}^K W_{ijk} v_{ik} h_j - \sum_{i=1}^M \sum_{k=1}^K b_{ik} v_{ik} - \sum_{j=1}^F a_j h_j \quad (5.1)$$

Probability Function:

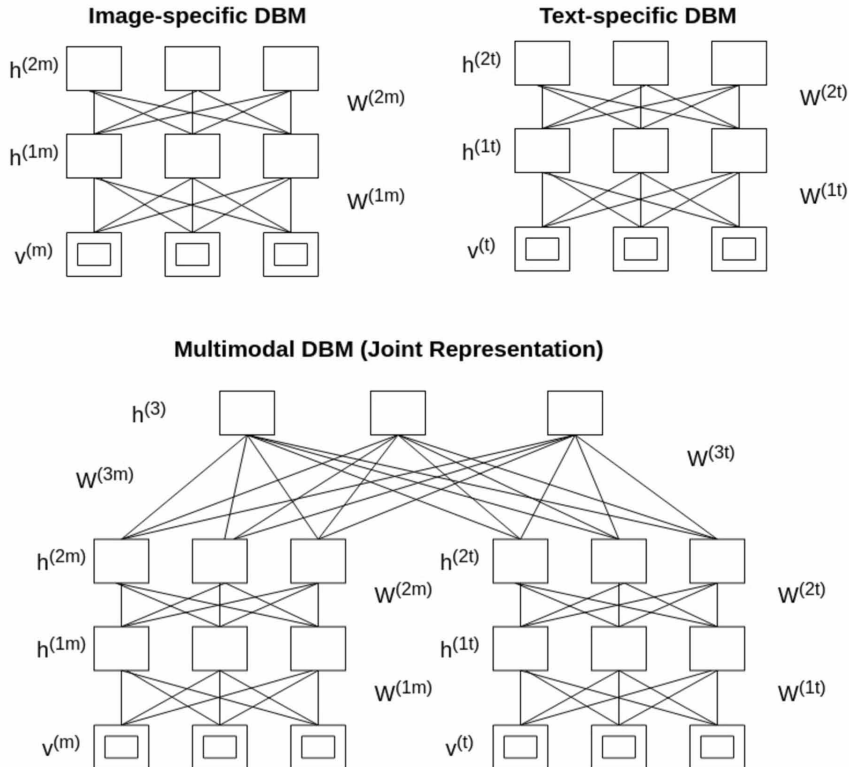
$$P(V, h; \theta) = \frac{1}{Z(\theta)} \exp(-E(V, h; \theta)), Z(\theta) = \sum_V \sum_h \exp(-E(V, h; \theta)) \quad (5.2)$$

A **Deep Boltzmann Machine** is a network of symmetrically coupled binary units and it contains the set of visible and hidden 3 layered DBM:

$$E(v, h; \theta) = -\sum_{i=1}^D \sum_{j=1}^{F_1} W_{ij}^{(1)} v_i h_j^{(1)} - \sum_{j=1}^{F_1} \sum_{l=1}^{F_2} W_{jl}^{(2)} h_j^{(1)} h_l^{(2)} - \sum_{l=1}^{F_2} \sum_{p=1}^{F_3} W_{lp}^{(3)} h_l^{(2)} h_p^{(3)} \\ - \sum_{i=1}^D b_i v_i - \sum_{j=1}^{F_1} b_j^{(1)} h_j^{(1)} - \sum_{l=1}^{F_2} b_l^{(2)} h_l^{(2)} - \sum_{p=1}^{F_3} b_p^{(3)} h_p^{(3)} \quad (6)$$

And so when we use **Multimodal DBM**, it is far more advanced version of the current RBM and uses the below architecture infused with Softmax Replication and Gaussian-Bernoulli RBM.

Figure 2. Image Specific: Gaussian-Bernoulli RBM and Text Specific: Softmax Replication RBM and joining both of them Multimodal DBM



Multimodal RBM belongs to the family of unsupervised artificial neural network. Probability function of the above figure can be given as:

$$\begin{aligned}
 P(v^m, v^t; \theta) &= \sum_{h^{(2m)}, h^{(2t)}, h^{(3)}} P(h^{(2m)}, h^{(2t)}, h^{(3)}) \left(\sum_{h^{(1m)}} P(v_m, h^{(1m)} | h^{(2m)}) \right) \left(\sum_{h^{(1t)}} P(v^t, h^{(1t)} | h^{(2t)}) \right) \\
 &= \frac{1}{Z_M(\theta)} \sum_h \exp \left(\underbrace{\sum_{kj} W_{kj}^{(1t)} v_k^t h_j^{(1t)} + \sum_{jl} W_{jl}^{(2t)} h_j^{(1t)} h_l^{(2t)} + \sum_k b_k^t v_k^t + M \sum_j b_j^{(1t)} h_j^{(1t)} + \sum_l b_l^{(2t)} h_l^{(2t)}}_{\text{Re plicated_softmax_text_pathway}} \right. \\
 &\quad \left. - \sum_i \frac{(v_i^m - b_i^m)^2}{2\sigma_i^2} + \sum_{ij} \frac{v_i^m}{\sigma_i} W_{ij}^{(1m)} h_j^{(1m)} + \sum_{jl} W_{jl}^{(2m)} h_j^{(1m)} h_l^{(2m)} + \sum_j b_j^{(1m)} h_j^{(1m)} + \sum_l b_l^{(2m)} h_l^{(2m)} \right. \\
 &\quad \left. + \underbrace{\sum_{lp} W_{lp}^{(3t)} h_l^{(2t)} h_p^{(3)} + \sum_{lp} W_{lp}^{(3m)} h_l^{(2m)} h_p^{(3)} + \sum_p b_p^{(3)} h_p^{(3)}}_{\text{Joint_3rd_Layer}} \right)
 \end{aligned} \tag{7}$$

But DBMs started failing because, the model started showing lesser relevance measure than the supervised models and this model lacked scalability. Model size increased rapidly thereby increasing the error. Recurrent Neural Nets proved to be efficient in word vector transformation.

Recurrent Neural Nets (RNN)

Due to their capability for processing arbitrary-length sequences, recurrent neural nets are a natural choice of sequence modelling tasks. Recurrent neural nets create count vectorizer and pair the words, so that they can be used get the desired image. They create vector representations and using the previous knowledge of RNNs as discussed in previous section, they map it to the CNN (convolutional neural net) for image mapping. Earlier state-of-art techniques were used to create word representations like making use of concepts like tf-idf vectorizer, count vectorizer, lemmatizer, bag of words. But now this all work is done using RNN.

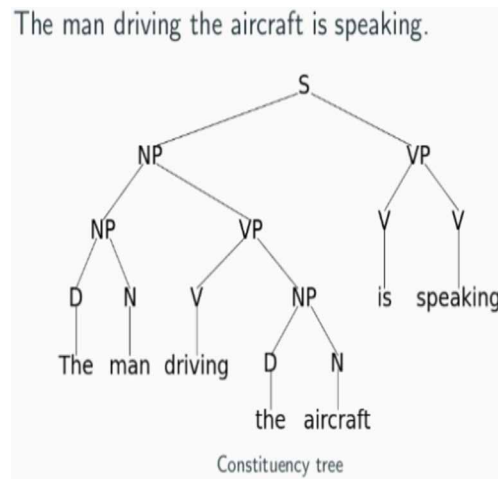
CT-RNN (Constituency Tree RNN)

Constituency Tree Recurrent Neural Net proves to be more efficient than the existing RNN architecture, certain function is used to compute the sentence representation, it's not like the simple RNN which averages the overall vector and maps the image based on that. CT-RNN use the binary tree data structures. They have the child-parent principle. These constituency tree RNNs (CT-RNNs) use the following composition function to compute a hidden parent vector 'h' from exactly two child vectors (c_1, c_2) in a binary tree:

$$h = f(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}) \tag{8}$$

where $W \in \mathbb{R}^{d \times 2d}$ is the main parameter to learn.

Figure 3.



DT-RNN (Dependency Tree RNN)

Sentence vectors of previous models cannot accurately represent visually grounded meaning. This led to the discovery of the DT-RNN model which uses dependency trees to embed sentences into a vector space in order to retrieve images that are described by those sentences. In this model, the network learns to map sentences (based on RNNs) and images into a common embedding space in order to be able to retrieve one from the other. Word and image representations are first learnt in their respective single modalities and then finally mapped into a jointly learned multimodal embedding space.

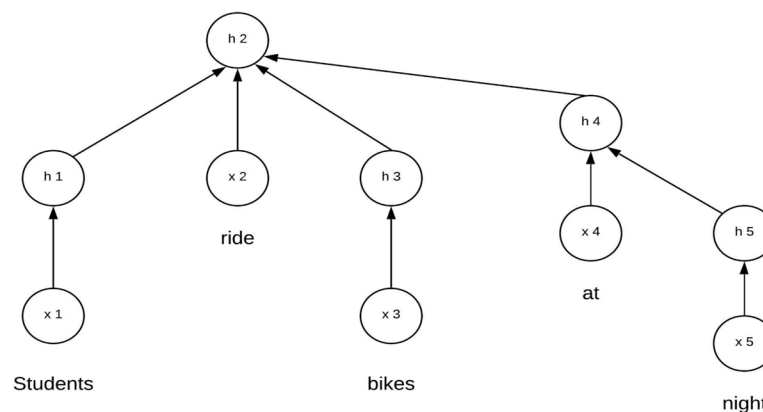
But learning word representations is not simply based on RNNs or CT-RNNs, there is a new architecture called Dependency Tree- RNN. The DT-RNN can solve the following problem:

The mother hugged her child.

The child was hugged by its mother.

Both the above sentences should map to roughly the same visual space but present recursive and recurrent neural networks do not exhibit this behavior and even bag of words representations would be influenced by the words 'was' and 'by'. The example, $\text{Students}_{(x1)} \text{ride}_{(x2)} \text{bikes}_{(x3)} \text{at}_{(x4)} \text{night}_{(x5)}$ computes sentence representation in bottom up fashion.

Figure 4.



In a bottom up fashion, the DT-RNN model computes parent vectors at each word that include all the dependent (children) nodes using a compositionality function g_θ which is parameterized by all the model parameters θ .

$$h_c = g_\theta(x_c) = f(W_v x_c) \quad (9)$$

Assume $l(i)$ be the number of leaf nodes (words) under node i and $C(i, y)$ be the set of child nodes of node i in dependency tree y . The final composition function for a node vector h_i becomes

$$h_i = f\left(\frac{1}{l(i)}(W_v x_i + \sum_{j \in C(i)} l(j) W_{pos(i,j)} h_j)\right) \quad (10)$$

where by definition $l(i) = 1 + \sum_{j \in C(i)} l(j)$ and $pos(i, j)$ is the relative position of child j with respect to node i . Further to add-on, semantic DT-RNN (SDT-RNN) is similar to the above equation except that the matrices $W_{pos(i,j)}$ are replaced with matrices based on the dependency relationship.

Long Short Term Memory (LSTM)

LSTMs have been discussed in the previous section, they emerge as an effective and representational power at capturing long term dependencies. There are different variations of LSTM, like:

Bidirectional LSTM, it consists of two LSTMs that are run in parallel: one on the input sequence and the other on the reverse of the input sequence. At each time step, the hidden state of the Bidirectional LSTM is the concatenation of the forward and backward hidden states. This setup allows the hidden state to capture both past and future information.

Multilayer LSTM, in the architecture, the hidden state of an LSTM unit in layer l is used as input to the LSTM unit in layer $l+1$ in the same time step. Here, the idea is to let the higher layers capture longer term dependencies of the input sequence.

Constituency Tree LSTM and Dependency Tree LSTM use the same above described functions to create word representations. Dependency Tree LSTMs prove to be most effective in compared to all the architectures discussed till now.

Dependency Tree structures prove to be useful and so few changes made in them and they create a whole new set of deep architectures. In an extensive research held by the authors, it is concluded that segmentation and fragmenting of images and word sentences have increased the accuracy in spite of the old architectures.

Attention Mechanism in LSTM-trees

Despite the fact that LSTMs take care of the RNN's vanishing gradient issue, they are constrained to linear data. LSTM these days' use tree-structured organization to experience the issue of dominant input words. On the off chance that basic LSTMs were utilized, at that point the last piece of the sentence gets more significance with respect to the starting part. But, if the tree structure of the sentence gets traversed from base to top then the data from various constituents initially gets combined to represent the root at the upper level and afterward this roots step by step gets traversed as children and joined to represent the root

at next dimension, etc. So in the two cases a LSTM cell will overlook past data which for plain LSTM, is identified with the length of the sentence and for Tree-LSTM, is identified with the depth of the tree.

When viewing the tree as a semantic representation of a sentence, this may not be the situation in numerous situations. For a constituency tree, if a sub-tree contains some negative sentiment words, at that point it isn't always the case that the constituent is negative. If the negative sentiment word is preceded by an invalidation (i.e. not bad, not ugly), at that point the constituent turns out to be semantically positive due to the nullification word (not). To catch this sort of data, attention is applied over the sub-tree parts to distribute the significance of each sub-tree segment when building the whole tree either semantically or linguistically. In the recent discovery, semantic attention is applied over the sub-tree components. Some attention weights are given to the children of the sub-tree and then the resultant is computed based on the weights.

SOLUTIONS AND RECOMMENDATIONS

Previous attempts to encode the word embeddings in LSTMs were only successful for smaller paragraphs and attention mechanism as discussed in the LSTM-tree architecture can be applied to all of the above. Multimodal and agent based learning as used in Deep Boltzmann Machines can be used similarly in other deep learning architectures as well. Mapping of image into vector space is also a crucial task as sentences with similar meanings might not end up in the same visual space which creates the room for error, these mistakes can be avoided by properly regulating the architectures using error functions and hyperparameter tuning. Certain type of generative network has come into existence which generates new data based on the previous data, and nowadays images don't come with tags, so this algorithm might come in handy.

FUTURE RESEARCH DIRECTIONS

Vectorized representation of the word sentences plays an important role in text based image retrieval. As discussed in the previous section there is an open window for using GANs (general adversarial networks) in order to overcome the problem of tagging which images in this era face and using attention mechanism on GANs may improve the performance, text based retrieval is a never ending field, some variations on the input side and functions involved will lead to the creation of new architecture. Further advancements have been made using image as the search query and also combining both (image and text as query) has proven to be quite successful. Some of the famous search engines also use image based information retrieval and the result comes out to be more promising than the past researches.

CONCLUSION

Deep learning has been proven to be a boon to the field of information retrieval and deep learning architectures have experienced a major breakthrough going from simple CNNs then to DBNs and then to RBMs as discussed in the previous version of the book. The chapter was concentrated only to the text based image retrieval but in the field of information retrieval itself there are different types which remain untouched. This article pours deep into the future of the chapter called, "An Insight to Deep

Architectures”, published in the encyclopedia. The innovative methodology of using RNNs in text based query processing which revolutionized NLP. This article also focused on the depth of the meaning of the sentence, how each and every word links with each other in order to get more accurate image results.

REFERENCES

Ahmed, M., Samee, M. R., & Mercer, R. (2019). *Improving Tree-LSTM with Tree Attention*. arXiv preprint arXiv:1901.00066

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., & Darrell, T. (2016). Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4555-4564). IEEE.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137). 10.1109/CVPR.2015.7298932

Karpathy, A., Joulin, A., & Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems* (pp. 1889-1897). Academic Press.

Lai & Hockenmaier. (2014). Illinoislh: A denotational and distributional approach to semantics. *SemEval 2014*.

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). *A structured self-attentive sentence embedding*. arXiv preprint arXiv:1703.03130

Mishra, A., Alahari, K., & Jawahar, C. V. (2013). Image retrieval using textual cues. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3040-3047). Academic Press.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., ... Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(4), 694–707.

Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2018). Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5706-5715). IEEE.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2, 207–218. doi:10.1162/tacl_a_00177

Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222-2230). Academic Press.

Tai, K. S., Socher, R., & Manning, C. D. (2015). *Improved semantic representations from tree-structured long short-term memory networks*. arXiv preprint arXiv:1503.00075

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1316-1324). IEEE.

Zhang, Y., Liu, Q., & Song, L. (2018). *Sentence-state lstm for text representation*. arXiv preprint arXiv:1805.02474

ADDITIONAL READING

Baqapuri, A. I. (2015). Deep learning applied to image and text matching. *arXiv preprint arXiv:1601.03478*.

Garg, N. P. N., & Tripathy, B. K. (2018). An Insight Into Deep Learning Architectures. In M. Khosrow-Pour, D.B.A. (Ed.), *Encyclopedia of Information Science and Technology*, Fourth Edition (pp. 4528-4534). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-2255-3.ch393

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Sequence Modeling: Recurrent and Recursive Nets. In *Deep learning*. MIT press.

Liu, W., Liu, P., Yang, Y., Gao, Y., & Yi, J. (2017). An Attention-Based Syntax-Tree and Tree-LSTM Model for Sentence Summarization. *International Journal of Performability Engineering*, 13(5).

Sumathy, P., Shanmugavadivu, P., & Vadivel, A. (2018). Text-Based Image Retrieval. In *Image Retrieval and Analysis Using Text and Fuzzy Shape Features: Emerging Research and Opportunities* (pp. 1–38). IGI Global., doi:10.4018/978-1-5225-3796-0.ch001

White, L., Togneri, R., Liu, W., & Bennamoun, M. (2019). Sentence Representations and Beyond. In *Neural Representations of Natural Language. Studies in Computational Intelligence* (Vol. 783). Springer. doi:10.1007/978-981-13-0062-2_5

KEY TERMS AND DEFINITIONS

Convolutional Neural Networks: A multi-layer neural network similar to artificial neural networks only differs in its architecture and mainly built to recognize visual patterns from image pixels.

Deep Boltzmann Machine: A deep Boltzmann machine (DBM) is a type of binary pairwise Markov random field (undirected probabilistic graphical model) with multiple layers of hidden random variables.

General Adversarial Network: General adversarial network (GAN) is a deep learning, unsupervised machine learning technique. It has a generator and a discriminator. Generator generates the new data and discriminator discriminates between generated input and the existing input so that to rectify the output.

Image Retrieval: An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most common utilized methods are adding metadata to captioning, keywords etc.

Long Short-Term Memory: Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

Recurrent Neural Networks: A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior.

Dynamic Load Balancing Using Honey Bee Algorithm: Load Balancing

Sudha S. Senthilkumar

Vellore Institute of Technology, India

Brindha K.

Vellore Institute of Technology, India

Nitesh Kumar Agrawal

Vellore Institute of Technology, India

Akshat Vaidya

Vellore Institute of Technology, India

INTRODUCTION

Distributed system include a set of interconnected and virtual computers that are provisioning dynamically. The Distributed system and cloud computing characteristics are basically as on-demand access, broad network access, resource pooling, rapid elasticity, measured service and security.

Providing the virtualization technology in cloud platforms helps enterprises to rent computing power in the form of virtual machines to the users. The users may utilize hundreds or thousands of Virtual Machines (VMs). To handle a very large size of data many techniques to optimize load and simplify operations are needed to obtain desirable performance level for the users. There is a need an effective algorithm in term of load balancing in the cloud computing. The load can be differentiated in various categories, CPU load, delay or network load, memory capacity. Load balancing guarantee that each processor in the system does approximately the same quantity of load at any point of time. The task scheduling is used to determine the suitable resources to execute the tasks that received from users. For increasing the utilization of resources, we use the load balancing algorithm. Efficient utilization is achieved by using the idle resources until finishing the resources of processors that have the high load. The load balancing mechanism divides the loads within all the resources which are available. The dynamic cloud computing used many of load balancing strategies. The Max-Min algorithm for each task, the minimum completion time is calculated, then the task with the maximum of minimum completion time is mapped to the corresponding virtual machine. The Shortest Job-First (SJF) is a various algorithm to CPU scheduling. This way associated with every process length's next CPU burst. The process with shortest length assigned to CPU when it is available. When there are two processes that have same CPU bursts then scheduling based on FCFS. In Round Robin (RR) approach, the processes are separated between all processor. Every process is mapping the processor in a round robin order by using quantum time. The allocation of process is not based on the allocations from remote processors. The processing time of task for various processes is not equal but the distributions of workload are same.

DOI: 10.4018/978-1-7998-3479-3.ch008

With the ever increasing size of the systems, there is a greater need for load balancing. The servers tend to get over worked and are rendered useless in catering to the needs of millions of users, and hence deny the services. So, a challenge in this modern era of smart computing is to make use of the various Load Balancing techniques to get the servers to function efficiently. Various algorithms are used for balancing the overall load of the cloud and a few of them are the honey bee foraging algorithm, a biased random sampling on a random walk procedure and Active clustering. Here, we will focus on the honeybee foraging algorithm and write a pseudo code for it, as well as develop a generic system model. The honeybee foraging algorithm is inspired from the behavior of honeybees for finding and reaping food. There is a type of bees called the forager bees who continually search for food sources and upon finding the same they return to the hive and advertise their discovery by a dance called as the waggle. The type of the movements in the dance reveal the quality and quantity of the food source as well as the distance of it from the beehive. Scout bees follow the forager bees to the location and reap the food. In case of load balancing in the web servers, whenever the demand sees a spike there is a dynamic allocation of services to regulate the changing demands of the user. The servers are grouped under Virtual servers (VS), each virtual server is assigned a specific queue for itself. Each server while processing a request calculates the reward and this is analogous to the quality of the find. The dance floor in case of the bees can be analogous to the advert board here which advertises the reward to the entire colony. Each server takes the role of either a forager or a scout. If the profit is high then the server stays at the current virtual server; posting an advertisement for its profitability (pr), if it was low then the server returns to the forage or scout behavior.

As mentioned earliest, the users send the requests to the service provider, the service provider has to serve many users to provide the best results and take into account maximizing resources utilization. Distributing the tasks into the resources to avoid overloading where there are under loaded resources. The objectives of this work to achieve users' satisfaction and provider satisfaction by balancing the loads and increase resources utilization, by minimizing the average of both make span time and response time using LBDA algorithm effective performance.

Proposed honey bee behavior inspired load balancing (HBB-LB) to achieve a good balance load through VMs for maximizing the throughput. We have compared the results of our proposed algorithm with the currently used algorithms. The results prove that our proposed algorithm has yielded better and efficient scheduling.

BACKGROUND

In an appropriated framework, dynamic load adjusting should be possible in two distinctive ways: distributed and non-distributed. In the conveyed one, the dynamic load adjusting calculation is executed by all hubs display in the framework and the errand of load adjusting is shared among them. The collaboration among hubs to accomplish stack adjusting can take two structures: cooperative and non-cooperative. Karaboga, D., & Ozturk, C. (2011)

In the first, the hubs work one next to the other to accomplish a typical target, for instance, to enhance the general reaction time, and so on. In the second frame, every hub works freely toward an objective neighborhood to it, for instance, to enhance the reaction time of a nearby local task. Dynamic load adjusting calculations of conveyed nature, normally create a larger number of messages than the non-dispersed ones on the grounds that, each of the hubs in the framework needs to connect with each other hub. An advantage, of this is regardless of the possibility that at least one hubs in the framework come

up short, it won't bring about the aggregate load adjusting procedure to stop, it's rather would impact the framework execution to some degree. Appropriated dynamic load adjusting can present massive weight on a framework in which every hub needs to trade status data with each other hub in the framework. It is more invaluable when a large portion of the hubs demonstration separately with not very many communications with others. Banerjee, S et al., (2009).

In non-distributed sort, it is possible that one hub or a gathering of hubs do the undertaking of load adjusting. Non-distributed dynamic load adjusting calculations can take two structures: centralized and semi-distributed. In the principal shape, the heap adjusting calculation is executed just by a solitary hub in the entire framework: the focal hub. This hub is exclusively in charge of load adjusting of the entire framework. Alternate hubs cooperate just with the focal hub. In semi-distributed frame, hubs of the framework are divided into bunches, where the heap adjusting in each group is of concentrated shape. George, G., & Raimond, K. (2013).

A focal hub is chosen in each group by suitable race procedure which deals with load adjusting inside that bunch. Subsequently, the heap adjusting of the entire framework is done by means of the focal hubs of each bunch.

Centralized dynamic load balancing takes less messages to achieve a choice, as the quantity of general associations in the framework diminishes radically when contrasted with the semi-distributed case. In any case, brought together calculations can bring about a bottleneck in the framework at the focal hub and furthermore the load balancing procedure is rendered pointless once the focal hub crashes. Accordingly, this calculation is most suited for systems with smaller size.

FOCUS OF THE ARTICLE

The enhanced artificial honey bee colony calculation (ABC), an advancement calculation in light of the savvy searching conduct of honey bee swarm was proposed by Karaboga, D., & Ozturk, C. (2011)

This metaheuristic is roused by the keen scrounging conduct of bumble bee swarm. The calculation exhibited in the work is for numerical function advancement. The advantage of ABC is that the global search ability in the algorithm is implemented by introducing neighbourhood source production mechanism Rastkhadiv, F., & Zamanifar, K. (2016).

In the ABC calculation, the honey bees in a settlement are partitioned into three gatherings: employed honey bees (forager honey bees), onlooker honey bees (observer honey bees) and scouts (explorer honey bees). The principal half of the settlement comprises of the employed honey bees and the second half incorporates the onlooker bees.

The employed honey bee whose nourishment source has been depleted by the honey bees turns into a scout for looking new sustenance sources haphazardly. An employed honey bee conveys the data about nourishment source and imparts this data to a specific likelihood by waggle move. Then again onlooker bees watch the waggle move as are put on the nourishment sources by utilizing a probability based choice process. As the nectar measure of a sustenance source builds, the probability with which the nourishment source is favored by onlooker bees increments, as well Tang, K., Zhang, J. M et al.,(2011).

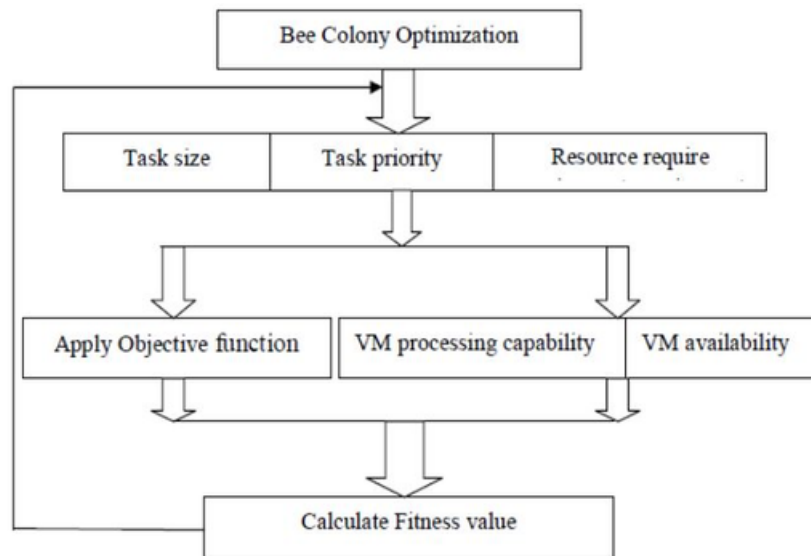
Bees can learn the directions and distances of their travels between hive and food sources by a process called path integration. A natural nourishment source is indicated in the honey bee's memory by a vector encoding distance and direction from the hive. This vector that is the yield of the way path integration process is employed for route on ensuing trips to the nourishment, and it is additionally what the honey

bee encodes in its waggle move. Some honey bee state based calculations have been planned with motivation from honey bee's navigational framework.

Calculations that take their motivation from the honey bee's home site choice conduct apply the guideline of choice making process that empowers a settlement to distinguish and unite towards one best arrangement Padhy, R. P., & Rao, P. (2011).

SOLUTIONS AND RECOMMENDATIONS

Figure 1. System Architecture



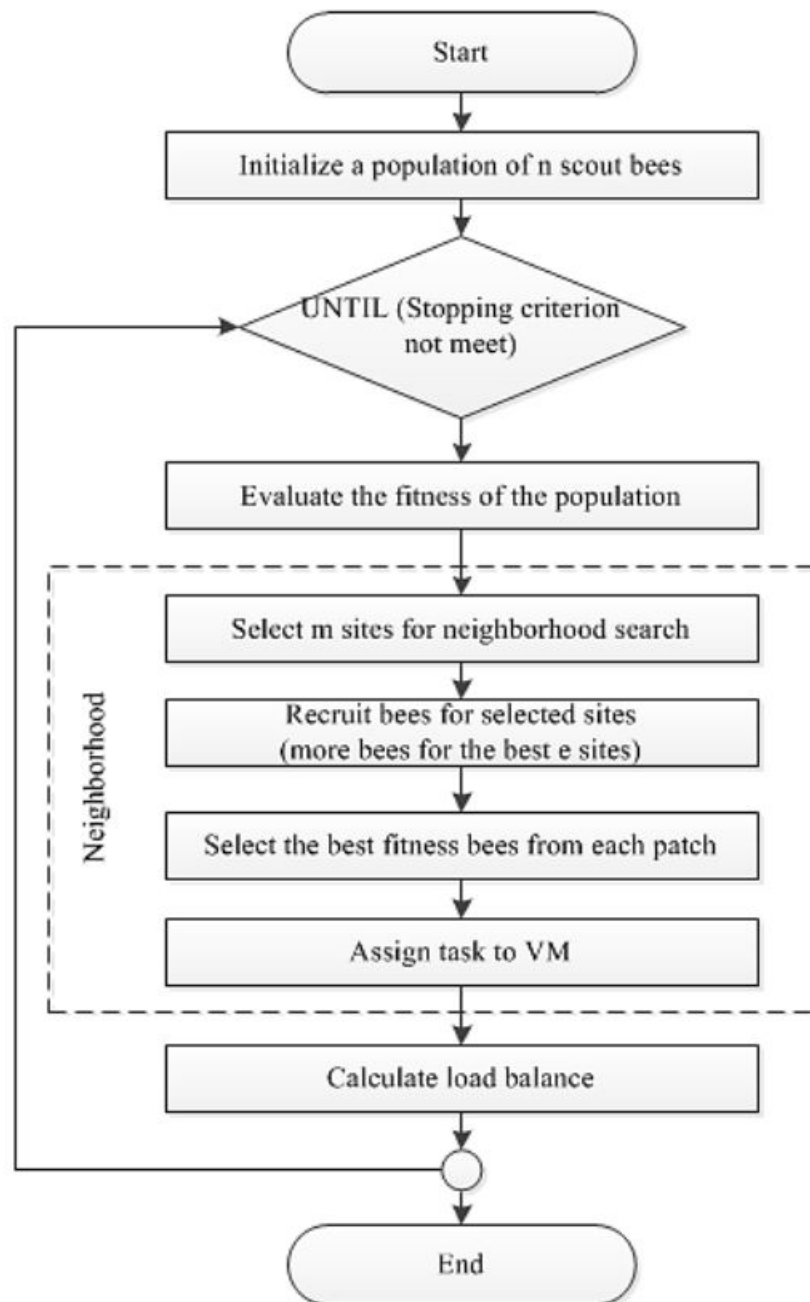
A users send a request about required resources, task priority and task size request to cloud manager in the bee colony optimization. The cloud Manager Store all requests. Now if forms a request table with required resources, task priority and task size request from all users. Then apply bee-Colony algorithm for all entry and Schedule and Assigns the channel according to Bee-Colony algorithms output Fitness value.

Honey Bee Behavior Inspired Load Balancing

1. Initialize a population of n scout bees.
2. Evaluate the fitness of the population.
3. Select the m sites for neighborhood search.
4. Recruit bees for selected sites.
5. Select the best fitness bees from each patch
6. Assign task to VM
7. Calculate load balance.
8. Stop the procedure

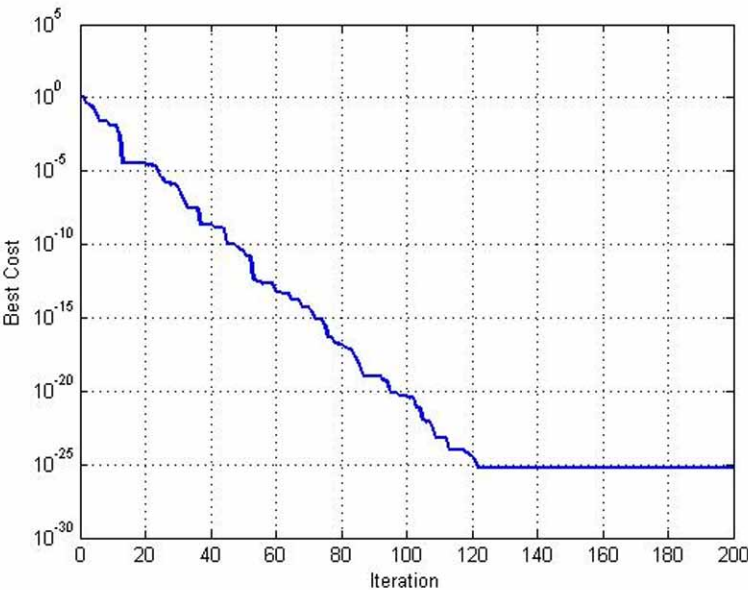
We implemented the Artificial Bee Colony Optimization algorithm to the numeric problem.

Figure 2. Flowchart for Honey bee behavior inspired load balancing (HBB-LB)



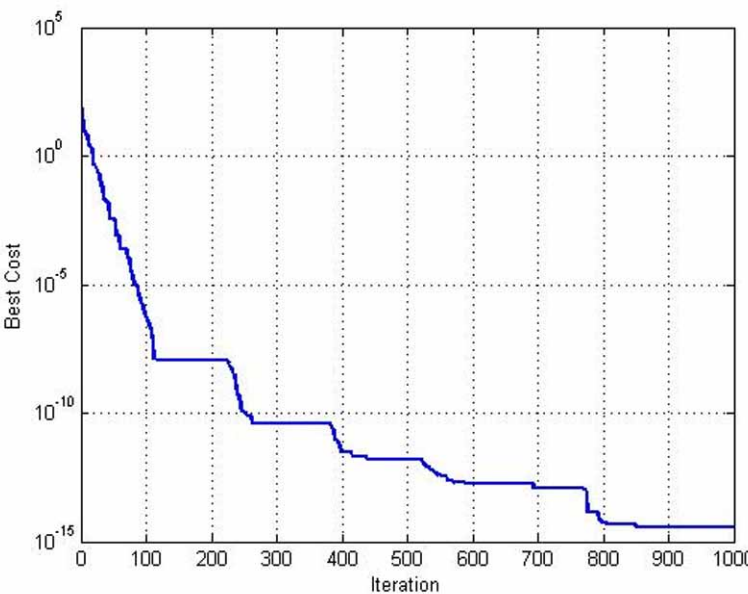
Test case	1
Population size	100
Iterations	200
Decision Variables	3
Range	[-5,5]

Figure 3. Iteration Vs Best Cost



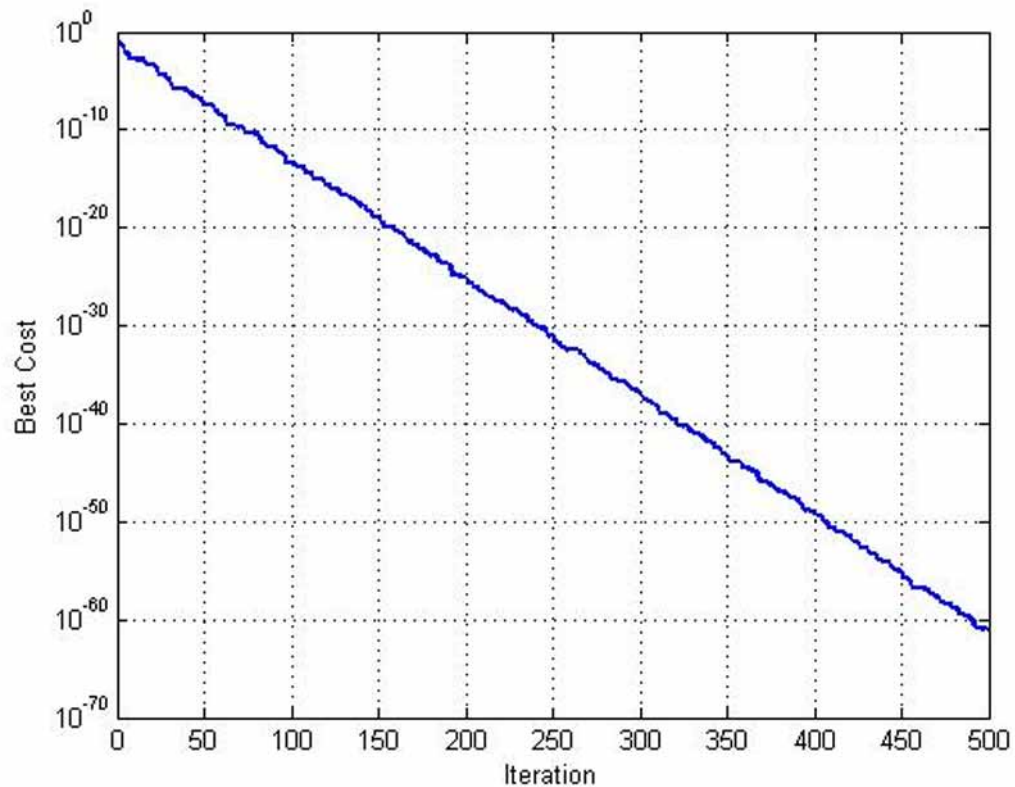
Test case	2
Population size	50
Iterations	1000
Decisions Variables	7
Range	[-10,10]

Figure 4. Iteration Vs Best Cost



Test case	3
Population Size	500
Iterations	500
Decision Variables	5
Range	[-1,1]

Figure 5. Iteration Vs Best Cost



FUTURE RESEARCH DIRECTIONS

There are various other aspects that could not be taken into consideration while researching for this work and can be further bettered and presented in future works.

CONCLUSION

Hence, upon observing a plethora of test cases, we can safely find the best case. Hence, there must be a ratio of number of different kinds of bees for us to reach at a better solution. Thus, we have shown three of the test cases out of the various that we have executed to show the general trend of the graphs upon changing the various test conditions.

REFERENCES

- Banerjee, S., Mukherjee, I., & Mahanti, P. K. (2009). Cloud computing initiative using modified ant colony framework. *World Academy of Science, Engineering and Technology*, 56(32), 221–224.
- George, G., & Raimond, K. (2013). A survey on optimization algorithms for optimizing the numerical functions. *International Journal of Computers and Applications*, 61(6).
- Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing*, 11(1), 652–657. doi:10.1016/j.asoc.2009.12.025
- Khajeh-Hosseini, A., Greenwood, D., & Sommerville, I. (2010, July). Cloud migration: A case study of migrating an enterprise it system to iaas. In *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on* (pp. 450–457). IEEE. 10.1109/CLOUD.2010.37
- Kruekaew, B., & Kimpan, W. (2014, March). Virtual machine scheduling management on cloud computing using artificial bee colony. In *Proceedings of the International MultiConference of engineers and computer scientists (Vol. 1, pp. 12–14)*. Academic Press.
- Pham, D. T., Ghanbarzadeh, A., Koç, E., Otri, S., Rahim, S., & Zaidi, M. (2006). The Bees Algorithm—A Novel Tool for Complex Optimisation Problems. In *Intelligent production machines and systems* (pp. 454–459). Academic Press.
- Rastkhadiv, F., & Zamanifar, K. (2016). Task scheduling based on load balancing using artificial bee colony in cloud computing environment. *International Journal of Advanced Biotechnology and Research*, 7(5).
- Sindhu, S., & Mukherjee, S. (2011). Efficient task scheduling algorithms for cloud computing environment. In *High Performance Architecture and Grid Computing* (pp. 79–83). Springer. doi:10.1007/978-3-642-22577-2_11
- Tang, K., Zhang, J. M., & Feng, C. H. (2011, October). Application centric lifecycle framework in cloud. In *e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on* (pp. 329–334). IEEE. 10.1109/ICEBE.2011.32
- Wang, S. C., Yan, K. Q., Liao, W. P., & Wang, S. S. (2010, July). Towards a load balancing in a three-level cloud computing network. In *Computer Science and information technology (ICCSIT), 2010 3rd IEEE International Conference on* (Vol. 1, pp. 108–113). IEEE.

ADDITIONAL READING

- Agarwal, A., Manisha, G., Milind, R. N., & Shylaja, S. S. (2014, April). A Survey of Cloud Based Load Balancing Techniques. In *Proceedings of International Conference on Electrical, Electronics, Computer Science and Mechanical Engg* (pp. 9–13).
- Aruna, M., Bhanu, D., & Punithagowri, R. (2013). A survey on load balancing algorithms in cloud environment. *International Journal of Computers and Applications*, 82(16).
- Gangwar, I., & Rana, P. (2014). Juxtaposition of load balancing algorithms in cloud computing using cloud analyst simulator. *International Journal of Computers and Applications*, 97(2).

- Ghomi, E. J., Rahmani, A. M., & Qader, N. N. (2017). Load-balancing algorithms in cloud computing: A survey. *Journal of Network and Computer Applications*, 88, 50–71. doi:10.1016/j.jnca.2017.04.007
- kaur, J., & King, S. (2012). A survey on load balancing techniques in cloud computing. *International Journal of Science and Research (IJSR) ISSN (Online)*, 2319-7064.
- Kumar, R., & Sahoo, G. (2013). Load balancing using ant colony in cloud computing. *arXiv preprint arXiv:1312.2074*.
- Mishra, R., & Jaiswal, A. (2012). Ant colony optimization: A solution of load balancing in cloud. *International Journal of Web & Semantic Technology*, 3(2), 33–50. doi:10.5121/ijwest.2012.3203
- Padhy, R. P., & Rao, P. (2011). *Load balancing in cloud computing systems* (Doctoral dissertation).
- Uma, J., Ramasamy, V., & Kaleeswaran, A. (2014). Load Balancing Algorithms in Cloud Computing Environment-A Methodical Comparison. *International Journal of Engineering Research & Technology (Ahmedabad)*, 3(2), 272–275.

KEY TERMS AND DEFINITIONS

Artificial Honeybee Colony Calculation (ABC): It is a swarm based meta-heuristic algorithm that was introduced by Karaboga in 2005 for optimizing numerical problems.

Dynamic Load Balancing: Assignment of tasks to a computational platform, with the objective of balancing the work assigned to processors. When the tasks are interdependent an additional objective is to minimize communication between processors. The load balance is monitored during execution of the application, and the assignment is modified if required, while trying to minimize the communication cost of migrating tasks.

Honeybee Algorithm: The bees algorithm is a method of problem solving that mimics the behavior of honeybees to find the optimum solution.

Virtual Servers: A virtual server is a server that shares hardware and software resources with other operating systems (OS) versus dedicated servers.

Waggle Dance: A wagging movement performed by a honeybee at the hive or nest, to indicate to other bees the direction and distance of a source of food.

Unsupervised Automatic Keyphrases Extraction on Italian Datasets

1

Isabella Gagliardi

IMATI-CNR, Italy

Maria Teresa Artese

IMATI-CNR, Italy

INTRODUCTION

Keyphrase extraction is an important research activity in text mining, natural language processing, and information retrieval. Keyphrases provide a compact semantic representation of the content of a document. Many tools for text management that automate tasks requiring high skill and expertise can benefit from the keyword extraction process (Zhang, 2008). To obtain the keywords, the texts should be analyzed to extract terms that characterize or represent the content of the documents and are useful in identifying documents relevant to a given query and/or in “suggesting” something in some way related. Keyphrases should represent the content of a document in all its aspects and be general enough to represent more than a single item, as well as specific not to represent the whole set of items. Keyphrases should, therefore, satisfy the following characteristics: i) have good coverage; ii) be relevant; iii) be consistent and iv) be up-to-date.

This process can be performed manually (associated or identified by experts, for example museum curators, in the case of Cultural Heritage) or automatically. The manual association of keywords to documents requires time, skills, specialized staff, and more steps to ensure that the chosen terms are consistent, adequate, relevant, with a good coverage, sufficiently general and timely. A large number of algorithms, divided into supervised or unsupervised methods, have been designed and developed to solve the problem of automatic keyphrases extraction. Supervised methods typically consider this problem as a classification task, in which a model is trained on annotated data to determine whether a given phrase is a keyphrase or not. Supervised approaches and systems can be found in Turney, 2000; Hulth, 2003; Hulth & Megyesi, 2006; Zhang, 2006; Nguyen & Kan, 2007; Hong & Zhen, 2012. These systems require a great number of training data, which can be unrealistic in the web scenario, and yet show a bias towards the domain on which they are trained. Keyphrase extraction task in unsupervised approaches is considered as a ranking problem. Unsupervised means that no human supervision is required: the algorithms are able to identify autonomously, and without any prior training, the terms to be extracted.

The aim of the chapter is to critically discuss the Unsupervised Automatic Keyphrases Extraction algorithms, analyzing in depth their characteristics. The methods presented will be tested on different datasets, presenting in detail the data, the algorithms and the different options tested in the executions. Moreover, most of the studies and experiments have been conducted on texts in English, while there are few experiments concerning other languages, such as Italian. Particular attention will be paid to the evaluation of the results of the methods in two different languages, English and Italian.

The chapter is organized as follows: after the description of the state-of-the-art of unsupervised keyword extraction methods, the algorithms, the data, the pre-processing methods applied, and the

DOI: 10.4018/978-1-7998-3479-3.ch009

experimental runs are presented. The results obtained are then presented, compared and evaluated, and future research discussed.

BACKGROUND

To identify the most relevant keywords for a text, the following pipeline has to be performed, that mimics the Information Retrieval one:

- Pre-process data
- Apply Unsupervised Automatic Keyword Extraction Algorithms:
- Extract a list of candidate keywords /keyphrases using some heuristics,
- Score each candidate keywords/keyphrases, according to different criteria and methods,
- Select the first m keywords/keyphrases.
- Evaluate the results.

Each step will be described in detail below.

Pre-Processing

Pre-processing (Kannan & Gurusamy, 2014; Vijayarani, Ilamathi, & Nithya, 2015, Uysal, & Gunal, 2014) has the aim of preparing the texts for the algorithms of keyword extractions. In this chapter datasets in English and Italian languages are considered. The steps in both languages are the same, but the methods, tools, experiences available for the two languages are different. English-language datasets can rely on well-established pre-processing tools and methods capable of eliminating, if useful, stoplist words and terms belonging to defined grammatical categories. Tokenization, lemmatization/stemming, POS (part-of-speech) tagging tools, and almost standard stopword lists can be easily identified and applied to obtain acceptable results.

The situation for the Italian language is more nuanced: the Italian grammar, morphology, and syntax are more complicated. As an example, some part of speech, as nouns and adjectives, are variable, that is, they are modified according to the number (singular and plural) and gender (feminine and masculine). The adjective beautiful, for the positive grade, takes shapes ‘*bello*’ (masculine/singular), ‘*bella*’ (feminine/singular), ‘*belli*’ (masculine/plural), and ‘*belle*’ (feminine/plural).

For each piece of text, the following steps, rather standard, will be taken, not necessarily in this order:

1. **tokenization**, that is, division of the text into individual single/multi words;
2. annotation, that may include **POS tagging**;
3. **normalization**: lemmatization/stemming;
4. removal of the **stopwords** and/or specific grammatical categories.

Tokenization is the process of decomposing a text, considered as a continuous set of words – or a string-, into a set of terms, composed of a single or compound words. This apparently trivial process is language dependent, as the characters that indicate the end of the word are different from language to language. In Italian, for example, the apostrophe is a character of division of words. It is obligatory in case of elision, such as a *bell'amico* (good friend) or *quest'alunna* (this (girl) student). In these cases, two

distinct words should be identified. In English the use of the apostrophe is different, e.g. for the Saxon genitive, or in case of contraction or abbreviation, not giving rise to two different words.

Moreover, regardless of language, different methods can be applied to identify compound words. For example, you can use statistical methods that identify groups of words frequently together and consider them as n-gram, or use headlines of Wikipedia page as a dictionary or identify their presence in WordNet.

POS (part-of-speech) tagging is the process of associating the corresponding grammatical category to every single word. Table 1 shows possible grammatical categories, taken from “Italian tagset used in the TreeTagger¹”.

Table 1. Italian tagset used in the TreeTagger parameter file

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt
(Copyright Prof. Achim Stein, University of Stuttgart)
ABR abbreviation
ADJ adjective
ADV adverb
CON conjunction
DET:def definite article
DET:indef indefinite article
FW foreign word
INT interjection
LS list symbol
NOM noun
NPR name
NUM numeral
PON punctuation
PRE preposition
PRE:det preposition+article
PRO pronoun
PRO:demo demonstrative pronoun
PRO:indef indefinite pronoun
...
VER:cimp verb conjunctive imperfect
VER:cond verb conditional
VER:cpre verb conjunctive present
...

This process requires understanding the structure of the sentence because, in Italian as in English, the same word can belong to different grammatical categories (e.g. be an adjective, a noun or a verb) depending on the composition of the sentence. For example in Italian, ‘*danza*’ (dance) in ‘*lui danza*’ (he dances) is a verb, while ‘*la danza*’ (the dance) is a noun.

Normalization: the process of lemmatization/stemming reduces the words to the lemma or stem. The two processes are different. The Porter algorithm is the most widespread stemmer for the English language. For example, according to the Porter stemmer (Porter, 1980) or Snowball algorithm (Porter, 2001), ceremony, as ceremonies, become *ceremoni*. The lemmatizes, on the other hand, should bring the various inflected forms to a single lemma, that is, the voice present in the vocabulary. For example ceremony, ceremonies should lead to ceremony; or listening, (he) listens, (they) listened, ... to “to listen”. For Italian, there are no similar “universal” algorithms.

Stopword: traditionally, in information retrieval systems, texts to be indexed are provided without stopwords, those words that are so frequent that they have no meaning: for example and, in, is/are, Obviously, stoplists are different for Italian and English. Also, depending on the context of use, they

should be customized. There are various lists of generic stopwords, both for Italian and English. For example, in Ranks NL website² there are several stoplists for English, while only one for Italian³.

The results of the pre-processing are lists of headwords or vocabularies, to be used - together with the original data - for the next steps of text analysis.

Algorithms

Although the problem of extracting keyphrases able to represent the content of pieces of text in natural language has been faced since the first Information Retrieval systems appeared, the advent of new tools and techniques makes it still very actual: in literature, many papers are devoted to this topic (Hasan & Ng, 2010; Hasan & Ng, 2014; Merrouni, Frikh, & Ouhbi, 2016; Siddiqi & Sharan, 2015) and several approaches have been defined (Gollapalli, Caragea, Li, & Giles, 2015; Hulth, 2003; Mihalcea & Tarau, 2004; Rose, Engel, Cramer, & Cowley, 2010).

Keyphrase extraction task in unsupervised approaches can be gathered into statistical-based, graph-based and cluster-based approaches. In statistical-based approaches, texts are usually represented as matrices in which the statistical techniques are applied to rank the words by using Tf-idf term weighting metric (Manning, Raghavan, & Schütze, 2010). In graph-based methods each document is represented as a graph where vertices or nodes represent words, and edges are connected based on either lexical or semantic relations, such as a co-occurrence relation. Examples are TextRank (Mihalcea et al., 2004), RAKE (Rose et al., 2010), SingleRank or CollabRank (Wan & Xiao, 2008a, 2008b). Cluster-based methods extract terms, group them into clusters based on their semantic relatedness using Wikipedia and/or other co-occurrence similarity measures, and select phrases that contain one or more cluster centroids. Examples are KeyCluster (Liu, Li, Zheng, & Sun, 2009) or SemCluster (Alrehamy & Walker, 2017).

In the following, the methods used in these experiments are presented, that is *Tf-idf*, *TextRank*, *RAKE*. Besides the results obtained applying *Latent Semantic Indexing* and *Latent Dirichlet Allocation*, able to extract topics from texts, are also described and discussed.

Tf-idf

Tf-idf is the short form for *term frequency–inverse document frequency*, widely used in Information Retrieval (Salton & Buckley, 1988; Manning et al., 2010) to measure the importance of a word in a document, within a collection or corpus. The Tf-idf weight is composed by two terms: normalized Term Frequency (TF), and the Inverse Document Frequency (IDF), and computed as follows:

TF: Term Frequency measures how frequently a term occurs in a document. For normalization purpose, the term frequency is divided by the document length:

$$TF(td) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

IDF: Inverse Document Frequency measures how important a term is. Frequent terms are weighed down, while scaled up the rare ones, by computing the following:

$$IDF(t) = \log_e \frac{\text{Total number of documents}}{\text{Number of documents containing term } t}$$

Tf-idf assigns to a term a weight in a document that is: a) highest when the term occurs many times within a small number of documents (thus lending high discriminating power to those documents); b) lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal); and c) lowest when the term occurs in almost all documents.

Tf-idf has been used extensively in information retrieval systems, and in general in search engines because i) is easy to compute, ii) cosine ranking formula sorts documents by degree of similarity to the query. On the other hand, being based on the bag-of-words (BoW) model, it does not capture the position in the text, semantics, co-occurrences in different documents.

TextRank

TextRank is a graph-based ranking model for text processing (Mihalcea et al., 2004), is based on the same assumptions of Google's PageRank (Brin & Page, 2012; Page, Brin, Motwani, & Winograd, 1999), that is: i) important pages are linked by important pages, and ii) the PageRank value of a page is essentially based on the probability that a user visits that page.

TextRank models a text document as a graph where each word (or sentence, according to granularity chosen) in the document is represented as a node, and the semantic relation between each node is represented by an edge. A scoring algorithm will then be executed on the graph to assess the value of each word, which will be used to rank them, then top ranking words are selected as keywords. Keywords that are adjacent in the document are combined to form multi-word keywords.

Main characteristics of TextRank are: it

- finds keywords that appear “central” (like PageRank);
- can be applied to any text without prior training;
- is domain independent;
- works on each single document;
- can generate keyphrases.

RAKE

RAKE, short form for Rapid Automatic Keyword Extraction (Rose et al., 2010) is an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents. It is graph-based.

The algorithm is composed of the following steps: i) it extracts a set of candidate keywords (single words or sequences of contiguous words), ii) assigns a score to these candidate keywords, based on the co-occurrence graph: the score is based on the frequency of the words and their degree (the sum of lengths of all sequences where the word appears), iii) adjoins keywords and iv) extracts the top ranking keywords.

Main characteristics of RAKE are: it

- can be applied to any text without prior training;
- is domain independent;
- works on each single document;
- can generate more complicated phrases;
- needs a comprehensive and reliable list of stop words.

Latent Semantic Indexing (LSI)

The idea behind LSI is to compare words to find relevant documents, but what you really want is to compare the meanings or concepts of words. LSI (Dumais, 2004; Landauer, 2006) is a technique able to analyze relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

LSI assumes that words that are close in meaning will occur in similar pieces of text. A term-document matrix, like that created by Tf-idf model, is the starting point for the technique - called Singular Value Decomposition (SVD) - used to reduce the number of rows in the matrix, while preserving the similarity structure among columns, thus converting terms and documents to points in a lower-dimension space.

Steps to apply LSI to a set of documents:

- Create the word by document matrix. In this matrix, each index word is a row and each document is a column;
- Apply *Tf-idf* weighting to the matrix;
- Apply upon the matrix the SVD technique to obtain a reduced dimensional representation of matrix that emphasizes the strongest relationships and throws away the noise.
- Words are then compared by taking the cosine between the two vectors formed by any two rows. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words.

LSI is able to cluster documents into n topics, n defined ahead, assigning each topic to a set of words. The belonging of a document to a cluster is scored with a numeric value $[0,1]$. Depending on the number of documents and the homogeneity/variety within each document and between documents (intra and inter-document), the number of clusters can be difficult to define.

Latent Dirichlet Allocation (LDA)

LDA is a technique that automatically detects topics contained in documents (Blei, Ng, & Jordan, 2003). The steps to identify topics, words and to associate documents with topics are:

- define the number of topics: it can be done either using an informed estimate (for example the results of a previous analysis), simply by trial and error or based on the best Coherence and perplexity values calculated on different numbers of topics;
- the algorithm will assign each word to a temporary topic, semi-randomly. This means that if a word appears twice, this word can be assigned to different topics;
- (iterative) the algorithm will check and update topic assignments, looping through the words of each document. For each word, the topic assignment is updated according to two criteria: 1) How common is this word across topics? and 2) How prevalent are topics in the document?

The process of checking topic assignment is repeated for each word in every document, cycling through the entire collection of documents multiple times. This iterative updating is the key feature of the LDA to generate a final solution with coherent topics.

Both LSI and LDA are algorithms for topic modelling, and the result is a set of n topics, n defined ahead, each containing a set of scored words or phrases. According to the different implementations,

the number of words/phrase for each topic can be defined. Both the words/phrases assignment to topics and documents assignment to topics are scored. The choice of the number n of topics can be difficult, and methods other than trial and error can be implemented.

Evaluation

Designing evaluation metrics for automatic unsupervised keyphrase extraction algorithms is by no means an easy task. In general, the results are compared with the keywords chosen manually by experts or using gold standard datasets.

According to SemEval contexts⁴, in this chapter Recall and Precision, borrowed by Information Retrieval, will be used to assess the results.

Recall measures the ability of the system to retrieve *all* relevant items in the corpus and can be expressed as (see Table 2):

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{total number of relevant documents}}$$

Or, expressed according to confusion matrix (Table 3):

$$\text{Recall} = \frac{\text{True positive}}{\text{True Positive} + \text{False Negative}}$$

In other words, *recall* is the fraction of relevant documents that have been retrieved over the total amount of relevant documents and its value is 1 when all the relevant instances are retrieved, decreasing up to 0 when none relevant instance is retrieved.

Precision measures the ability of a search engine to retrieve *only* relevant items in the corpus and can be expressed as:

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{total number of retrieved documents}}$$

Or

$$\text{Precision} = \frac{\text{True positive}}{\text{True Positive} + \text{False Positive}}$$

Table 2. Classical Information Retrieval matrix

	Retrieved	not retrieved
Relevant	Retrieved & relevant	Retrieved & relevant
Not Relevant	Retrieved & not relevant	Not retrieved & not relevant

Table 3. Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision is the fraction of relevant documents among the retrieved documents, and its value varies from 1 to 0, with 1 if only relevant documents are returned, and 0 if none relevant document is retrieved. Sometimes it is used $P@n$, precision measured on the first n topmost results returned.

Figure 1 shows an example of computation of values of Recall, Precision and $P@n$. In an ideal system, both recall and precision should tend both to 1; in a real world system, depending on the user's needs, recall or precision tend to be favored: recall when there are few important documents to retrieve, precision when the dataset is wide, and the loss of some documents is less important than the retrieval of many useless documents.

Figure 1. Example of Recall and Precision computation

n	doc n.	relevant	
1	675	x	
2	690	x	
3	576		
4	590	x	
5	986		
6	592	x	
7	984		
8	988		
9	578		
10	985		
11	103		
12	591		
13	772	x	
14	990		

Let total number of relevant documents = 6

Recall = $1/6=0.167$; Precision = $1/1=1$

Recall = $2/6=0.333$; Precision = $2/2=1$

Recall = $3/6=0.5$; Precision = $3/4=0.75$

Recall = $4/6=0.667$; Precision = $4/6=0.667$

Recall = $5/6=0.83$; Precision = $5/13=0.38$

Missing one relevant document.
Never reach 100% recall

Adapted from Marco Gori slide on Evaluation of Multimedia Data Base class <http://www.dii.unisi.it/~marco/bdm/>

THE EXPERIMENTATION

In the following the experiment test bed will be described, starting from the datasets, the pre-processing methods applied, concluding with the algorithms and their evaluation, showing some results obtained.

Datasets

To test the algorithms 6 different datasets, in Italian and English languages, have been used, as described below:

CookIT

1

CookIT portal⁵ stores multimedia information describing, in Italian language, the traditional recipes handed down from generation to generation. The choice of the recipes to be inserted has been made by interviewing people of different ages, social conditions, and Italian regions to identify the recipes they consider as part of their own tradition and culture. Recipes are collected from the most relevant traditional recipes websites in Italy, such as Cucina Italiana⁶, Giallo Zafferano⁷. At the moment, 140 recipes have been added to the portal, and about 450 traditional Italian recipes have been identified and are being edited.

Lombardy Digital Archive: AESS (Archivio di Etnografia e Storia Sociale)

The AESS database (Artese & Gagliardi, 2012) stores information related to the oral history of the Lombardy region. The documents (interviews, songs, transcription of texts, transcription of music, scores, manuscripts, sound files, images, photographic documents, videos) are the result of targeted ethnographic and anthropological field research. They are an impressive and continuously updated corpus of traditional intangible culture in the Lombardy territories. Currently, the archive consists of 15,000 oral documents, among them about 4500 records related to photographs, in Italian language, have been used for the experimentation.

Intangible Search (Italian)

This is an online collection of “living good” of Lombardy Region and Alp territories, created from Lombardy Digital Archive (Artese & Gagliardi, 2013). IntangibleSearch⁸ offers information in several languages: Italian, English, French and German⁹, being the result of a cross-border EU project, ECHI, involving partners from all over the Alpine region: Italians, Swiss, French. 254 Italian-language documents, available on the web, have been used here.

Intangible Search (English)

This is the English subset of the IntangibleSearch, described above. About 166 English-language documents have been used.

Europeana

Europeana¹⁰ is the European platform for digital cultural heritage, built as a result of several EU-funded projects (Haslhofer & Isaac, 2011). More than 3,000 European institutions have contributed to building this vast collection, allowing users to explore Europe’s cultural and scientific heritage from prehistory to the present day. A set of metadata, in all European languages, is collected and integrated into a single collection, while data is stored in the original location, without any duplication. For this experiment, metadata has been extracted, through some queries, considering only those in English, for a total of 1928 documents.

Victoria and Albert Museum

The V&A museum¹¹ is the world's leading art and design museum, with a vast collection of over 2.3 million objects covering over 5,000 years of human creativity. The collections cover the theater, art books, painting, glass, ceramic architecture, furniture, fashion, textiles, photography, sculpture, jewelry, Asian art and design. Detailed information and descriptions have been extracted for this experiment, together with tags (if available), for a total of 558 documents.

Table 5 summarizes the main information on the 6 datasets used, three in Italian and three in English. If present, tags associated by catalogers or experts have been used in the tests.

Table 5. information on the 6 datasets used

	Cookit	AESS	Intangible	Intangible	Europeana	V&A
Language	Italian	Italian	Italian	English	English	English
Topic	Traditional Recipes	Intangible Cultural Heritage	Intangible Cultural heritage	Intangible Cultural Heritage	Cultural Heritage	Cultural Heritage
Doc. number	143	4771	264	166	1928	558
Average Sentence no.	10,60	1,85	12,71	15,65	3,16	9,89
Average Words no.	288	35	454	465	58	248
Tags no.	47	4588	263	165	682	539

With the aim of testing the algorithms, and identifying if (and, if so, which) method produces better results on this data with a set of test queries, and if there is any difference between English and Italian, a standard Information Retrieval approach was followed.

Pre-Processing

For the datasets in English and Italian, the pre-processing phases are the same, while the specific tools and algorithms to be applied differ. For the English language, data sets were processed using standard tools, such as Stanford's core NLP suite¹², Natural Language Toolkit of Python (Loper & Bird, 2004), or the NLTK package with PENN Treebank as a POS tagger and tokenizer (Marcus, Marcinkiewicz, & Santorini, 1993). Initially, the same tools were also used for Italian - where they were present - but the results obtained were not useful due to the presence of too many errors and inaccuracies, both for the POS tagging and for the lemmatization. Subsequently, tools developed specifically for the Italian language were tested, for example, the Italian version of Snowball¹³, Pattern python package specific for Italian (Pattern clips 2.6) or Lingua (Dell'Orletta, 2009; Attardi, Dell'Orletta, Simi, & Turian, 2009). In the end, TreeTagger (Schmid, 1994) was chosen, a free tool developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart, using the Italian standard tagset. A second tagset was also tested, but the results were not as satisfactory.

We applied different pre-processing methods to the datasets:

Table 6. The 10 top most frequent terms for the 2 datasets, according to the different pre-processing methods applied

The top 10 Most Frequent terms					
Dataset\pre-processing method	orig_nouns	orig_nouns_adj	tokens	tokens_nouns	tokens_nouns_adj
Cookit	minuto acqua olio impasto salalsale fuoco pasta forno farina modo	minuto acqua olio impasto salalsale fuoco pasta forno farina piatto	il minuto poi aggiungere fare acqua cuocere lasciare circa olio	acqua olio fuoco impasto pasta sala farina forno modo ciotola	minuto acqua olio fuoco impasto pasta sala forno farina piatto
Intangible English	people procession year time village group part church place use	people procession year small time village group part church traditional	use make also one take procession people place year two	use people year procession place time day village group man	use people year procession place time small day village group

Table 7. The 10 top least frequent terms for the 2 datasets, according to the different pre-processing methods applied

The top 10 Least Frequent terms					
Dataset\pre-processing method	orig_nouns	orig_nouns_adj	tokens	tokens_nouns	tokens_nouns_adj
Cookit	coda sfera pandolce entroterra aria caratteristica farcitura timballo seguito crostaceo	coda sfera pandolce entroterra aria sorrentina amarognolo saporito posto caratteristica	dissalare assorbimento coda sfera pandolce rispettare seguire sgusciare sorrentina vegetale	aosta assorbimento coda sfera pandolce orzo timballo crostaceo risultato lonza	assorbimento coda sfera pandolce contorno amarognolo orzo piemontese cernia timballo
Intangible English	selection worshipper row repetition zone homage dairy wreath sheep-shearing capetan	sheep-shearing humidity direct elegant established orchestra strap hospital suit lorry	wreath humidity elegant established settlement orchestra 100 badocher interpret Onofrio	humidity settlement orchestra badocher diameter straw strap army hospital strike lorry	wreath humidity elegant settlement orchestra badocher strap total army hospital

Original data: no pre-processing steps applied (orig)

Tokens: only tokenization of texts (tokens)

Tokens more than 1: only tokens appearing more than once are kept (tokens_moreonce)

After POS tagging, tokens that are nouns (`_nouns`), and tokens that are nouns and adjectives (`_nouns_adj`) were held on the three processed datasets (orig, tokens, tokens_moreonce).

The 10 Most frequent and the 10 Least frequent terms are shown in Table 5 and Table 6 for 2 datasets (one data set in Italian and one in English) according to the different pre-processing methods applied. You can see that the different pre-processing methods select a set of different terms, thus leading to very different keywords/keyphrases extracted by the algorithms.¹⁴

Algorithms

The following Unsupervised Keyphrases Extraction algorithms, described in section 2, have been tested and evaluated:

- Tf-idf
- TextRank
- RAKE
- Latent Semantic Indexing
- Latent Dirichlet Allocation.

Tf-idf was tested using 1-gram basis for word extraction: in Artese & Gagliardi, 2018 results are reported for n-gram tests; for TextRank only nouns and adjectives have been selected as potential keywords; for LSI, tests were performed with a different number of topics: here are reported the results for 20 and 100 topics, for LDA perplexity values and coherence were evaluated to identify the best number of topics.

Some algorithms have to be performed document by document (TextRank, RAKE), while others must be executed on the whole collection of documents (Tf-idf, LSI, LDA).

According to the Information Retrieval diagram, automatically extracted keyphrases are then used for document indexing in a spatial vector model and cosine similarity is the metric chosen to retrieve and rank documents in response to the request. Each execution on an algorithm/dataset/pre-processing method gives rise to a different set of keywords (weighted), associated with each document: so that the list of retrieved documents can be different both in the documents and in the retrieval order. Table 7 shows information and some statistics on the results obtained, applying the 5 algorithms on Italian Cookit pre-processed dataset. For other datasets/pre-processing methods, similar values have been obtained. In this phase, the same parameters were applied to the pre-processed datasets, having modified, as described, stopword lists, POS tagging, and stemming algorithms.

Evaluation

The implemented algorithms were tested on the (pre-processed) datasets, using the queries of which those in Table 8 are a subset. The results were assessed in terms of recall and precision.

TextRank and RAKE index each document only with top ranked terms and when a query is executed, the matching algorithm retrieves and scores only those documents that have terms in common with the query, the other documents are not retrieved. On the other hand tf-idf and (to some extent) LSI and LDA

Table 8. Some information and statistics on the results obtained on the Cookit dataset

Algorithm/ Dataset	Pre-processing method	Average text length	Average number of keyphrases per doc	No. different keyphrases	Average length of keyphrases	No. keyphrases with more than 10 occurrences
Tf-idf						
Cookit	orig	1822	19.81	2134	1	5
	orig_nouns	621	19.97	1597	1	12
	orig nounsadj	821	20	1899	1	7
	tokens	1411	20	1834	1	7
	tokens nouns	509	19.88	1367	1	22
	tokens nounsadj	749	20	1792	1	8
TextRank						
Cookit	orig	1822	5.58	427	1.36	9
	orig_nouns	621	5.5	510	1.7	7
	orig nounsadj	821	7.07	621	1.7	9
	tokens	1411	5.9	446	1.49	9
	tokens nouns	509	5.07	454	1.74	5
	tokens nounsadj	749	7.25	625	1.79	550
RAKE				s		
Cookit	orig.	1822	10.8	695	1.159	18
	orig_nouns	621	-	1		
	orig nounsadj	821	-	7	1.42	
	tokens	1411	-	10	1.1	
	tokens nouns	509	-	2		
	tokens nounsadj	749	-	2		
LSI 20 topics						
Cookit		1822	10	74	1	43
	orig_nouns	621	10	67	1	39
	orig nounsadj	821	10	67	1	41
	tokens	1411	10	68	1	39
	tokens nouns	509	10	62	1	33
	tokens nounsadj	749	10	60	1	30
LSI 100 topics						
Cookit	orig	1822	9.6	187	1	51
	orig_nouns	621	9.9	168	1	41
	orig nounsadj	821	9.8	178	1	50
	tokens	1411	10	177	1	49
	tokens nouns	509	10	171	1	35
	tokens nounsadj	749	10	177	1	51
LDA						
Cookit	orig	1822	10	112	1	37
	orig_nouns	621	10	102	1	31
	orig nounsadj	821	10	105	1	33
	tokens	1411	10	109	1	36
	tokens nouns	509	10	100	1	30
	tokens nounsadj	749	10	106	1	33

Table 9. Example of queries used to test the algorithms

Cookit	AESS	Intangible-IT	Intangible-EN	Europeana	V&A
impasto zucchero forno mozzarella mandorla spada acciuga	piadena anno impianto fotografia cascina festa lavoro	pietra processione latte chiesa statua carnevale paese	statue mask puppet procession carnival church people	silk lace plate needle cake credit brussels	miniature oil portrait ivory simple glass watercolour

algorithms give a weight to each term in the document: thus, in essence, documents are sorted according to the weight of the terms in the query.

For this reason, the precision of TextRank and RAKE is always very high and around 1, thus confirming the high quality of the algorithm results. On the other hand, recall values can be low because only those documents to which the searched terms have been associated are retrieved.

Tf-idf, LSI and LDA order the results according to the cosine similarity values of the vector space. Tf-idf obtains precision@10 and P@20 value almost close to 1, with lower precision on 100 items. LSI and LDA have high value of recall, at the expenses of precision. The results of LSI vary as the number of records varies. For example, on AESS, which contains more than 4000 documents, the results of LSI 20 show that very different elements are grouped together in the same topic. So we can conclude that 20 clusters are too few. LDA algorithm uses coherence and perplexity values give hints on the optimal number of topics.

For example, Victoria & Albert museum dataset has been evaluated on ‘cheese’ query. The dataset contains 84 items to be retrieved.

With *Tokens* pre-processing, Tf-idf retrieves 74 items, RAKE 0, TextRank 14, LSI-100 302 and LDA 120 with a number of topics of 35. For Tf-idf precision value is 1, and recall is 0,88, for RAKE 0, while TextRank has precision 1 and recall 0,18, LSI-100 has recall 1,00 and precision 0,24, and LDA 35 has recall 0,89 and precision 0,61.

With *Orig* pre-processing Tf-idf retrieves 73 items, RAKE 8, TextRank 9, LSI-100 319 and LDA-35 139. For Tf-idf precision value is 1, and recall is 0,86, for RAKE 0, while TextRank has precision 1 and recall 0,18, LSI-100 has recall 1,00 and precision 0,24 and LDA-35 has recall 0,96 and precision 0,57.

On another query, ‘portrait’, the same dataset contains 270 items to be retrieved.

With *Tokens* pre-processing, Tf-idf retrieves 273 items (of which relevant 263), RAKE 2, TextRank 49 (of which relevant 41), LSI-100 404, and LDA-35 342. For Tf-idf precision value is 0,96, and recall is 0,97, for RAKE almost 0, while TextRank has precision 0,83 and recall 0,15, LSI-100 has recall 1,00 and precision 0,66. LDA-35 has recall 0,99 and precision 0,77.

With *Orig* pre-processing Tf-idf retrieves 264 items, RAKE 50, TextRank 41, LSI-100 406, and LDA-35 345. For Tf-idf precision value is 1, and recall is 0,86, for RAKE 0, while TextRank has precision 0,83 and recall 0,15, LSI-100 has recall 1,00 and precision 0,66. LDA-35 has recall 0,99 and precision 0,76.

The experimental setup has been implemented in Python 2.7, using standard packages like Numpy, Matplotlib, Pandas and other more specific ones for processing of textual data such as NLTK (Loper & Bird, 2004), Gensim (Řehůřek & Sojka, 2011), Newspaper¹⁵, Pattern (Pattern clips 2.6), and Sklearn¹⁶ together with some experimental packages in GitHub.

SOLUTIONS AND RECOMMENDATIONS

The algorithms for automatic unsupervised extraction of keywords/keyphrases are a topic of great scientific interest, given the large availability of texts on the web. The experiments described have shown that by changing methods, tools and applying different parameter values to the algorithms, the results can also vary greatly.

It is worth noting that:

- Tf-idf, TextRank and RAKE algorithms produce scored keywords, while LSI and LDA produce n topics composed of unscored keywords.
- RAKE obtains results only on the original texts, without the lemmatization and elimination of stopwords;
- the number of different keyphrases varies greatly on the basis of the algorithm tested and pre-processing applied: from 67 of LSI with 20 topics to over 2000 of tf-idf, excluding RAKE;
- TextRank produces a relatively large number of keywords, using the recommended values: setting a threshold value, as the maximum number of keywords extracted per document or the weight of the nodes must be further tested and evaluated;
- TextRank, RAKE extract keyphrases longer than 1.
- LSI basically works as a clustering algorithm, associating each cluster with a set of keywords that represent it. The results reported were obtained by choosing 20 and 100 as the number of topics.
- LDA too works as a clustering algorithm: to identify the optimal number of topics, per each dataset/pre-processing method, coherence and perplexity have been evaluated. For Cookit dataset the optimal number of topics is 25.

The analysis of these results shows that the algorithms behave in the same way on datasets in English and Italian. What would require further analysis is the Italian-language pre-processing. For example in pos tagging/lemming, all the systems tested show errors or inaccuracies in both processes, especially with verbs in the presence of combined pronouns, resulting in a faulty recognition.

FUTURE RESEARCH DIRECTIONS

Tests have shown that further work could lead to optimized results and should be geared towards:

- Identify and test other pre-processing tools, specifically designed for the Italian language for better POS tagging and lemmatization results;
- implement and test other algorithms and methods, such as Keycluster, SingleRank, ExpandRank, or deep learning methods;
- integrate different algorithms (results) to obtain the best keywords;
- integrate word2vec, GloVe, WordNet or Babelnet to manage synonyms and identify topics in a more efficient and effective way;
- extend the experimentation to other languages, such as French, German or Spanish;
- test algorithms on other media files and/or on larger data sets;

- further study evaluation tasks, both testing different ways to create a set of keywords for comparison/evaluation, and investigating other methods, borrowed also from information retrieval or image tagging.

CONCLUSION

The aims of these experiments have been to test and evaluate different automatic unsupervised keyword/keyphrase extraction algorithms; and evaluate if there are differences in processing Italian vs English datasets.

Five algorithms present in literature have been tested on six different datasets and the results of the experimentation have been reported here, to evaluate which algorithm works best, in terms of recall and precision, with respect to query tests. The experiments described have shown that by changing methods, tools and applying different parameter values to the algorithms, the results can also vary greatly.

The results on Italian vs English datasets, evaluated in terms of recall and precision, are mainly influenced by pre-processing methods and specific tools, rather than by algorithms.

REFERENCES

- Alrehamy, H. H., & Walker, C. (2017, September). SemCluster: unsupervised automatic keyphrase extraction using affinity propagation. In *UK Workshop on Computational Intelligence* (pp. 222-235). Springer.
- Artese, M. T., & Gagliardi, I. (2012, October). Cataloging intangible cultural heritage on the web. In *Euro-Mediterranean Conference* (pp. 676-683). Springer. 10.1007/978-3-642-34234-9_71
- Artese, M. T., & Gagliardi, I. (2013, October). Browsing and searching UNESCO Intangible heritage on the web: two ways. In *2013 Digital Heritage International Congress (DigitalHeritage)* (Vol. 2, pp. 417-420). IEEE. 10.1109/DigitalHeritage.2013.6744794
- Artese, M. T., & Gagliardi, I. (2018, June). What is this painting about? Experiments on Unsupervised Keyphrases Extraction algorithms. *IOP Conference Series. Materials Science and Engineering*, 364(1), 012050. doi:10.1088/1757-899X/364/1/012050
- Attardi, G., Dell'Orletta, F., Simi, M., & Turian, J. (2009). Accurate dependency parsing with a stacked multilayer perceptron. *Proceedings of EVALITA*, 9, 1–8.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18), 3825–3833. doi:10.1016/j.comnet.2012.10.007
- Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. *Proceedings of EVALITA*, 9, 1–8.
- Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science & Technology*, 38(1), 188–230. doi:10.1002/aris.1440380105
- Gollapalli, S. D., Caragea, C., Li, X., & Giles, C. L. (2015). *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*. ACL.

Hasan, K. S., & Ng, V. (2010, August). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 365-373). Association for Computational Linguistics.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1262-1273). Academic Press.

Haslhofer, B., & Isaac, A. (2011). *data. europeana. eu*-The Europeana Linked Open Data Pilot.

Hong, B., & Zhen, D. (2012). An extended keyword extraction method. *Physics Procedia*, 24, 1120–1127. doi:10.1016/j.phpro.2012.02.167

Hulth, A. (2003, July). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 216-223). ACL. 10.3115/1119355.1119383

Hulth, A., & Megyesi, B. B. (2006, July). A study on automatically extracted keywords in text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* (pp. 537-544). ACL. 10.3115/1220175.1220243

Kannan, D. S., & Gurusamy, V. (2014). Preprocessing Techniques for Text Mining. *International Journal of Computer Science & Communication Networks*, 5(1), 7–16.

Landauer, T. K. (2006). *Latent semantic analysis*. John Wiley & Sons, Ltd. doi:10.1002/0470018860.s00561

Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009, August). Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 257-266). ACL. 10.3115/1699510.1699544

Loper, E., & Bird, S. (2004). NLTK: the natural language toolkit. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100–103.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). *Building a large annotated corpus of English: The Penn Treebank*. Academic Press.

Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2016, October). Automatic keyphrase extraction: An overview of the state of the art. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)* (pp. 306-313). IEEE.

Mihalcea, R., & Tarau, P. (2004). Texttrank: Bringing order into text. *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Nguyen, T. D., & Kan, M. Y. (2007, December). Keyphrase extraction in scientific publications. In *International conference on Asian digital libraries* (pp. 317-326). Springer.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Pattern clips Software (Version 2.6) [Software]. A web mining module for the Python programming language. (n.d.). Available from <https://www.clips.uantwerpen.be/pattern>

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. Academic Press.

Řehůřek, R., & Sojka, P. (2011). *Gensim—Statistical Semantics in Python*. Academic Press.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1-20.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Schmid, H. (1994). Probabilistic pos tagging using decision trees. *Proceedings of International Conference on New methods in Language Processing*.

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computers and Applications*, 109(2).

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4), 303–336. doi:10.1023/A:1009976227802

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, 50(1), 104–112. doi:10.1016/j.ipm.2013.08.006

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7–16.

Wan, X., & Xiao, J. (2008, August). CollabRank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 969-976). Association for Computational Linguistics. 10.3115/1599081.1599203

Wan, X., & Xiao, J. (2008, July). *Single Document Keyphrase Extraction Using Neighborhood Knowledge* (Vol. 8). AAAI.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computer Information Systems*, 4(3), 1169–1180.

Zhang, K., Xu, H., Tang, J., & Li, J. (2006, June). Keyword extraction using support vector machine. In *International Conference on Web-Age Information Management* (pp. 85-96). Springer.

ADDITIONAL READING

Bharti, S. K., & Babu, K. S. (2017). Automatic keyword extraction for text summarization: A survey. *arXiv preprint arXiv:1704.03242*.

Budanitsky, A., & Hirst, G. (2001, June). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources* (Vol. 2, pp. 2-2).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Navigli, R., & Ponzetto, S. P. (2010, July). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 216-225). Association for Computational Linguistics.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250. doi:10.1016/j.artint.2012.07.001

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004* (pp. 38-41). Association for Computational Linguistics.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26(3), 13. doi:10.1145/1361684.1361686

KEY TERMS AND DEFINITIONS

Automatic Keyword Extraction Algorithms: Algorithms able to extract keywords from texts, automatically and in an unsupervised manner. They are usually based on probability, graph, or on clusters.

Bag of Documents Model: Is a representation of a text as a set of its words, holding multiplicity (how many times each word appears), but losing position and other related information.

Dataset: Is a collection of data. Usually referred to a coherent set of data, such as records of a single database, or even of a single table.

Keywords/Keyphrases: Single or compound words able to characterize the content of the documents, useful for identifying the documents relevant to a given query.

Information Retrieval: Is the activity of searching information, relevant to user's needs from a collection of documents. Documents usually are texts, but can be images, video, or sounds.

Lemmatization: Is the process of reducing inflected forms of a word to its dictionary form or lemma.

POS (Part of Speech) Tagging: Is the process of assigning to each word of a text a particular part-of-speech, as verb, nouns, adjective, adverb, and so on. Different languages need different POS tagging tools.

Stemming: Is the process of reducing inflected forms of a word to its root or stem.

ENDNOTES

¹ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

² <https://www.ranks.nl/stopwords>

³ <https://www.ranks.nl/stopwords/italian>
⁴ SemEval-2017, End-User Development using NL: <http://alt.qcri.org/semeval2017/task11/index.php?id=task-description>
⁵ <http://arm.mi.imati.cnr.it/cookIT/>
⁶ <https://www.lacucinaitaliana.it/ricette/>
⁷ <http://www.giallozafferano.it>
⁸ <http://intangibleSearch.eu>
⁹ reported in the same order as the website.
¹⁰ <http://europeana.eu>
¹¹ <https://www.vam.ac.uk/>
¹² <https://stanfordnlp.github.io/CoreNLP/>
¹³ <http://snowball.tartarus.org/algorithms/italian/stemmer.html>
¹⁴ When the work is completed, the datasets used will be made available on a shared platform, together with the test results and related statistics
¹⁵ <https://github.com/codelucas/newspaper>
¹⁶ Scikit-learn: machine learning in Python, <https://scikit-learn.org/>

An Intelligent Machine–Driven Perspective to Archaeological Pottery Reassembly

1

Wilson Sakpere <https://orcid.org/0000-0003-2429-8924>*University of Bologna, Italy***Valentina Gallerani***University of Bologna, Italy*

INTRODUCTION

Information is an imperative concept that humans use to communicate because communicating aids understanding. Within an information, there could be other components of data such as images, video, numbers and text. The need to stay informed in modern day has resulted in various ways of transmitting information with the use of technologies such as the computer and internet. Hence, Information and Communication Technologies (ICT) has played an important role in this aspect. ICT is the use of modern equipment and methods to transmit and exchange information. ICT has helped in understanding the human society and way of life. The applications of ICT include research and development in a variety of application fields such as teaching, transport, business, entertainment, politics, agriculture, medicine and various industry and academic disciplines among others. One of such disciplines, that this article is based on, is archaeology.

Archaeology is a discipline that recovers, studies and analyses ancient and modern artefacts and material remains from past societies of humans and closely related species, recording its primary data for the purpose of reusing them for historical understanding of human culture (Jeffrey, 2010:47). Archaeologists have used ICT as a means of making, keeping and preserving records and information to make it available for further research and future use (Eiteljorg, 2004; Marchetti *et al.*, 2017; A. Smith, 2004). This is to considerably reduce the likelihood of information degradation and loss (A. Smith, 2004). Hence, various methods have been used by social scientists and digital humanists to preserve historical artefacts and relics such as literary texts and arts, frescoes and potsherds (Sabharwal, 2015:60-61).

In particular, archaeologists have employed different time-consuming means of studying and preserving historical relics (Marchetti *et al.*, 2017). Such methods include physical examination, classification, illustration and reconstruction of potsherds through drawing. These methods are carried out manually, making the archaeologist spend time and effort to observe and possibly extrapolate from the observations (Drucker, 2012:89). However, reconstruction of many excavated fragments of potsherds is still a major concern. This paper proposes a path towards an optimal means to reconstruct excavated archaeological potsherds to a high standard. To achieve this, image analysis techniques (computer vision) as well as machine learning algorithms are considered.

In addition, the terms ‘reconstruct’ and ‘reassemble’ are used interchangeably in this paper to mean the same thing. The remainder of this paper is structured as follows: Section 2 reviews some related works from a general point of view and highlights the potentials inherent in computer vision and ma-

DOI: 10.4018/978-1-7998-3479-3.ch010

chine learning; Section 3 delves into the research problem, identifying the research questions, methods, objectives and methodology that this study is based on; Section 4 concludes our discussion.

BACKGROUND

As archaeologists study the historic and prehistoric human activities through reconstructing discovered ancient artefacts to have a grounded understanding of the prehistoric culture, the ability to carry out these reconstructions accurately and within reasonable time become pertinent. In this technology era of virtual and augmented reality, among others, researchers have been applying the use of digital technologies to make work faster and output optimal. This has helped in making professional life easier, more enjoyable and more productive, leading to what is now known as ‘smart environment’. Therefore, applying technological means to solve critical issues in the archaeological field is desirable.

Hence, many studies have attempted to solve the reconstruction problem. For example, the study of Kampel and Sablatnig (2003) developed a system that can process both complete and broken vessels. This was achieved using two reconstruction strategies known as: “shape from silhouette based method for complete vessels and a profile based method for fragments” (Kampel & Sablatnig, 2003). While using these strategies have improved performance with an acceptable accuracy, it is nonetheless dependent on certain conditions being met, thus requiring further investigations that will improve accuracy.

Likewise, Smith, Bespalov, Shokoufandeh and Jeppson (2010) investigated methods of classification and reconstruction of images of excavated archaeological ceramic fragments based on their colour and texture information. This was achieved by using Scale Invariant Feature Transform (SIFT) and a feature descriptor based on Total Variation Geometry (TVG) (P. Smith *et al.*, 2010). However, the performance of these descriptors was not satisfactory in terms of effectiveness and accuracy. Also, the study of Puglisi, Stanco, Barone and Mazzoleni (2015) proposed a system that uses SIFT as well. The system uses image processing techniques to automatically identify and analyse images of the structural components of potsherds through “optical microscopy”. As a result, the suitable features can then be computed and analysed for classification purposes. In addition, this system aims at segmenting the acquired images and extracting their features for pottery classification. To achieve this and for the sake of better accuracy, the authors chose to use SIFT feature point method, where the feature points are extracted and matched with related pairs. While this system improves on the ones before it, it falls short in providing an optimal solution with a high accuracy, hence opening the path to further studies (Puglisi *et al.*, 2015).

Furthermore, Makridis and Daras (2012) focused on a technique for accurate and automatic classification of potsherds. The technique was implemented in four steps namely: feature extraction, feature fusion, feature selection and classification. This approach attempted to reduce the computational complexity problem with a “bag of words (BoW)” method that uses the features extracted from the images to create a “global descriptor” vector for the images (Makridis & Daras, 2012). However, while this technique turned out to be quite efficient, how and if this solution reduces computational complexity is still an open issue.

Finally, Roman-Rangel, Jimenez-Badillo and Marchand-Maillet (2016) proposed Histogram of Spherical Orientations (HoSO), a method for automatic analysis of potsherds by applying computer vision techniques. This method analyses the external frontal view of the potsherd alone, processes 3D data by using only the points coordinates without using colour, texture or faces, and efficiently encodes the information from the points coordinates. They posit that the advantages include substantial time reduc-

tion in potsherd organisation and a simple method of acquiring the image (Roman-Rangel *et al.*, 2016). This shows that advantages are inherent in the application of computer vision techniques.

Potentials of Computer Vision and Machine Learning in Pottery

Among various modern digital techniques that are becoming relevant for archaeological studies, both computer vision and machine learning are emerging. Vision is the ability to use the brain to sense and interpret the surrounding environment with light reflected by objects in the environment. This is demonstrated when archaeologists apply the manual method of potsherds reassembly. On the other hand, computer vision is the science of building systems that make useful decisions to obtain information from digital images or complex image data (Forsyth & Ponce, 2012). In other words, the ability to make machines to see and do what the human perception can. Computer vision can perform feature detection, extraction, fusion and selection, then analyse it to understand the information perceived from images or imagery data (Forsyth & Ponce, 2012). Other potentials of computer vision include reconstruction, tracking, recognition, restoration and learning. In addition, a computer vision system that is well designed will increase productivity.

Image Acquisition and Processing

In any image processing project, the image acquisition stage is the first. Without this stage, there is no project that can be done. The image is first acquired by an optical device such as a camera and processed into a manageable form. According to Moeslund (2012), “the image acquisition process consists of three steps namely: energy reflected from the object of interest, an optical system which focuses the energy and finally a sensor which measures the amount of energy”. Also, various methods exist for acquiring the image depending on the project to be done. After the image has been acquired, processing and enhancements can be carried out. However, the image acquisition will need to be done in a satisfactory manner for the processing to be satisfactorily done.

While the image can be acquired in 2D and converted into 3D, it can also be directly acquired in 3D. In the studies of Díaz-Marín *et al.* (2015), Stamatopoulos and Anagnostopoulos (2016) and Tsiafaki, Koutsoudis, Arnaoutoglou and Michailidou (2016), image acquisition was effected in 3D. However, the methods of acquisition differ from one to the other. Where Díaz-Marín *et al.* and Tsiafaki *et al.* used a laser scanner and a camera to acquire images in 3D, Stamatopoulos and Anagnostopoulos acquired images in 2D by panoramic photography with a digital camera, which were then transformed into 3D with a “specialised software”.

Although image acquisition in 3D has many advantages over acquisition in 2D, its use and prevalence in cultural heritage requires more investigation. While 2D images give a limited information of the physical shape and size of an object, 3D images express the geometry in terms of three-dimensional coordinates. Furthermore, current technological advancement in camera and laser technologies have made the production of an optimal and accurate 3D depth data achievable. Hence, numerous 3D data acquisition systems have been developed. These systems can be applied in cultural heritage studies, and archaeology in particular, as is the case in this study.

Machine and Deep Learning

Machines are capable of learning and processing information around them, a process that can be referred to as pattern recognition and learning techniques. These techniques are usually categorised as artificial intelligence or machine learning. For example, supervised learning is a machine learning technique. While computer vision techniques have been extensively exploited in the attempt to automate pottery reconstruction, as cited above, no serious effort has been devoted to the application of new techniques based on machine learning.

Machine learning allows machines to accurately predict outcomes for the future by learning about the past without the machine being explicitly programmed (Haykin, 2009; Marsland, 2015). Machine learning algorithms are categorised into supervised learning and unsupervised learning. In supervised learning, both the input and the desired output are provided with supervision so that the machine can learn and apply the knowledge to new dataset (Alpaydm, 2010; Haykin, 2009). In unsupervised learning, only input data is provided unsupervised while the machine iteratively computes and arrives at its deduction. Machine learning has helped in proffering solutions to problems in various disciplines. While prominent results of application of these techniques include: self-driving cars, speech recognition, face recognition, biometric recognition, spam mails detection, effective web searches, applications in computer vision, robotics and medical diagnosis, and a host of other improvements (Alpaydm, 2010; Marsland, 2015), optimal performance could be obtained when these methods and techniques are applied to classification and reassembly problems. With these potentials in view, cultural heritage disciplines and archaeology in particular can benefit from these technologies. Hence, the considerations of computer vision and machine learning in the application of this study.

POTSHERD REASSEMBLY: A CASE STUDY

As an example of how image analysis techniques as well as machine learning algorithms could be useful in the archaeological context, how these techniques could be put to good use in the specific field of potsherds reassembly is here presented. This should be considered as a paradigmatic and advanced case application of ICT in archaeology, since documentation and reassembly of excavated potsherds is still one of the most pressing concerns in a typical field project (Marchetti *et al.*, 2017).

Of all the artefacts recovered from archaeological excavations, pottery is one of the most significant because of its ability not to decay when compared with other archaeological artefacts of other materials (Barclay *et al.*, 2016). Depending on how the pottery is preserved, it could be recovered as a complete pot on site or as fragments known as potsherds. Most potsherds, especially the ones of the pottery rim, usually carry information such as the shape of the pottery, decoration on the pottery and colour of the pottery among others, on the inner and/or outer surfaces (Di Angelo, Di Stefano, & Pane, 2017). This information helps the archaeologist to carry out studies such as dating, and to understand the culture of the potter. However, to be able to carry out a meaningful study, there is a need to reassemble these potsherds as much as possible for a more comprehensive understanding of what is being studied. Thus, the reconstruction problem is recognised as of primary importance. Figure 1 shows some excavated archaeological potsherds.

Furthermore, reconstruction is usually done manually by attempting to physically arrange and join the potsherds together just like in a puzzle. Going a step further, reconstruction can also be done by drawing the pottery form by extrapolation from the arranged or classified potsherds (Collett, 2012) as shown in

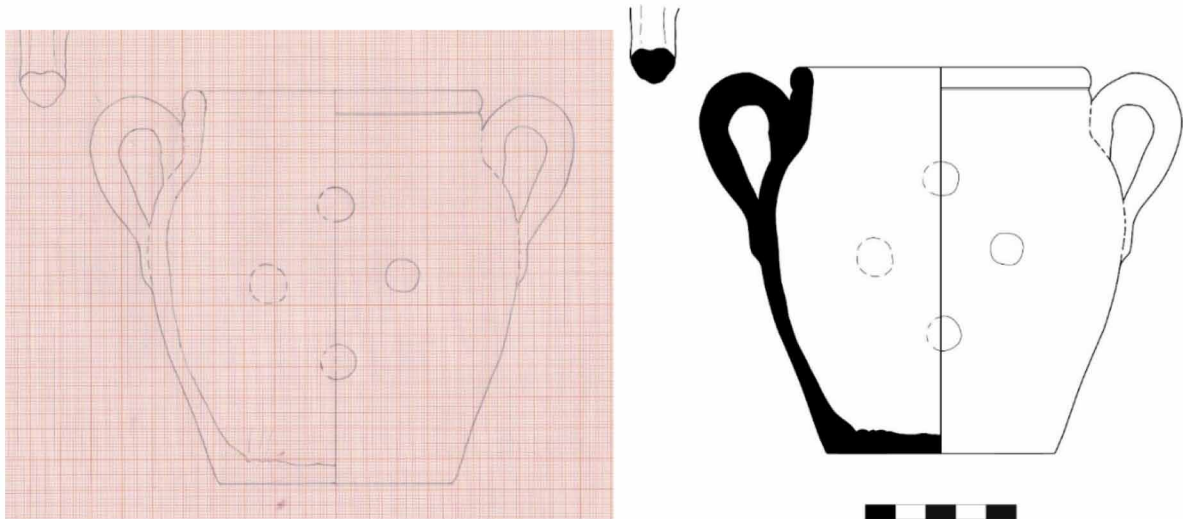
Figure 1. Sherds of excavated Archaeological Pottery (Source: Creative Commons¹)



Figure 2; a drawing of the sectional and external views of a modern pottery vase. They represent the first step of conventional drawing by pencil on paper through comb, other technical tools and its subsequent inking, which is then effected by drawing digitally on a raster image of the pencil drawing.

However, when certain basic tasks are not taken into consideration in the reconstruction process, there is the tendency to spend a long time trying to solve the puzzle of reconstruction and not achieving it, leading to a lack of understanding and missing valuable information (Barclay *et al.*, 2016). Thus,

Figure 2. A hand-drawn and Adobe Illustrator drawing of a modern pottery vase with the left side showing the section view and the right side showing the external view



requiring better ways of approaching this challenge.

PROBLEM STATEMENT

Excavated fragments of archaeological pottery have been quite challenging to reconstruct into desired artefacts (Colantoni & Gottarelli, 2008; Makridis & Daras, 2012). According to Colantoni and Gottarelli (2008), such challenge was observed in the process of classification and study of excavated potsherds, reconstruction of potsherds by physical assembly and reconstruction of potsherds by drawing. In addition,

potsherds recovered from archaeological sites are rarely complete enough to provide useful information, hence archaeologists fall back on the time-consuming method of reconstructive drawing by extrapolation (Di Angelo *et al.*, 2017; Drucker, 2012:89). Reconstructing potsherds is significant in the field of pottery and archaeological studies, analysis and research. Barclay *et al.* (2016) note that the excavated potsherds are used as a reliable tool for dating, and as evidence for ancient tradition, culture, civilization and technology. In fact, the use of technology and computer aided techniques have impacted positively on the studies on pottery and archaeology (Collett, 2012).

However, potsherds reconstruction coupled with the time, effort and cost expended, make an optimal means of reconstruction vital. According to Barclay *et al.*, “If the study of pottery is to reach its full potential, it is vital that it is recovered and analysed to a high standard.” Hence, being able to assemble and reconstruct potsherds optimally and with ease will greatly impact pottery and archaeological studies. Unfortunately, there is little evidence that various research works have addressed this persisting issue satisfactorily with respect to a high standard reconstruction of pottery (Marchetti *et al.*, 2017; Rice, 1987:286).

RESEARCH DESIGN AND METHODOLOGY

While computer-based attempts have been made to optimally reconstruct archaeological potsherds, a viable and optimal means of reconstruction is still desired. Hence the persisting gap for investigating the reconstruction of archaeological potsherds based on the contribution that new digital technologies, such as machine learning, can provide. Consequently, to address the gap, the following research questions in Table 1 are to be researched and responded to, with partial progress made on research questions one and two.

Research Approach

Images of the potsherds are first processed by an expert archaeologist, then processed digitally. The processed images are then acquired as training data for a semi-supervised learning-based machine. The system should then be able to infer thereafter and generate an output (pottery form) from the acquired data. From an epistemological viewpoint, such an approach falls under the realist and positivist approach,

Table 1. Research Problem, Research Questions and Corresponding Research Methods and Objectives

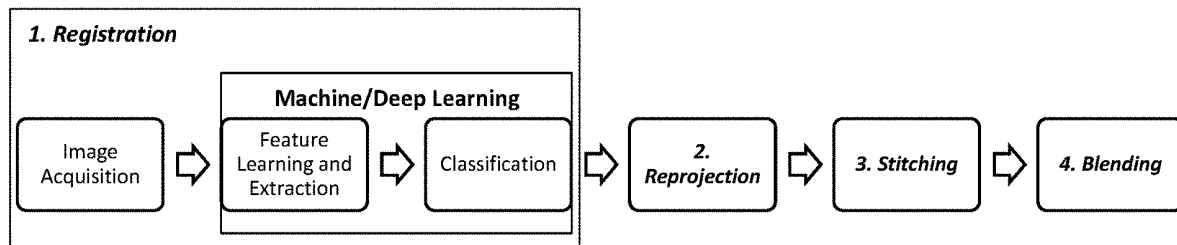
Research problem	Excavated fragments of archaeological potsherds are quite challenging to reconstruct to a high standard at present with current approaches, leading to unnecessary time consumption.	
Research Questions	Research Method(s)	Objectives
Why is it so difficult to digitally reconstruct archaeological potsherds?	Literature Analysis	To identify the challenges and drawbacks of techniques and methods in potsherd reconstruction.
How can archaeological potsherds be reconstructed to yield a high standard form for archaeological research purposes?	Literature Analysis and Experiment	Identify and determine best processes, techniques or technologies to use in implementing an optimal solution.
How should certain technologies be applied in the reconstruction process to enable a successful form assembly?	Literature Analysis and Experiment	Identify ways to implement a viable system, and develop a framework that would make potsherd reconstruction feasible.

where experiential methods and techniques of empirical enquiry prevail over abstract interpretations. On a more practical side, the objective of this methodology is to enhance and improve the accuracy of data capture and processing, leading to an optimal potsherd reconstruction through the experiential manipulation of the technical attributes or variables employed, while reducing time consumption and efforts expended by archaeologists. All this considered and due to the nature of this problem, the quantitative methodology of research, data collection and data analysis could be adopted. Hence the following strategy should be pursued as simply described in Figure 3.

Research Design and Method

Due to the nature of this study, data would be collected and analysed using computer vision techniques and a combination of machine learning and deep learning models. Particularly, image mosaicing methods will be employed. Because of the high accuracy of deep learning models, it will be considered for feature

Figure 3. Research Conceptual Model



learning and extraction. On the other hand, machine learning will be considered for classifying the data. Following the stages discussed in Ghosh and Kaabouch (2016), this experiment has been divided into the stages in Figure 3.

1. **Registration:** This refers to the establishment of geometric correspondence between a pair of images depicting the same scene. Registering a set of images entails an estimation of the geometric transformations that align the images with respect to a reference image within that set. The set may consist of two or more images taken of a single scene at different times, from different viewpoints, and/or by different sensors.
 - **Image acquisition:** While images of the potsherds that will be used as test data will be acquired in 2D and 3D, the 3D data will be considered in this study. A different set of labelled images of potsherds, and possibly textual and painted cultural heritage fragments will be acquired and archived as training data.
 - **Feature learning:** Here features of colours, edges, corners and anything that can represent the image will be extracted to facilitate pattern detection, recognition and classification – useful for prediction and identification. This will be done using deep neural network such as Convolutional Neural Network (CNN).
 - **Classification:** Classification is performed using the features obtained from learning. Appropriate classifiers will be used in MATLAB to identify the one that gives the best result. This will be done to determine shape characteristics, segmentation and other features required to differentiate between a class of potsherds from another class.

2. Reprojection: This refers to the alignment of the images into a common coordinate system using the computed geometric transformations. Appropriate methods are applied to find borders and regions that match.
3. Stitching: At this stage, geometric description will be determined, and fragments retrieved and matched. Stitching overlays the aligned images on a larger canvas by merging pixel values of the overlapping portions and retaining pixels where no overlap occurs. Errors propagated via geometric and photometric misalignments often result in undesirable object discontinuities and seam visibility around the boundary between two images.
4. Blending: The stitched images are cropped and seamlessly blended. Blending algorithm is used during or after the stitching step to minimize the discontinuities in the global appearance of the mosaic or assembled object.

FUTURE RESEARCH DIRECTIONS

Individuating an automatic machine learning approach to pottery reconstruction can be considered of interest to archaeologists, as they could then focus on other aspects of their studies, thereby increasing productivity. With the research questions one and two already partially explored, focus will thereafter be on the third research question as well. The anticipated outcome is a system that can optimally acquire potsherds digitally and reconstruct a pottery form with the available data. Where the data is not complete, i.e. some potsherds missing, the system should be able to extrapolate from available data to complete the reassembly of the pottery. While the solution is the goal, formulating a satisfactory process is also of great importance to this study. To this end, a framework able to make image processing a less complicated task needs to be explored, while the data collected from this activity will open further opportunities in this study.

CONCLUSION

This paper presents an approach towards reconstructing excavated archaeological potsherds to a high standard and with high accuracy. This approach involves the use and application of computer vision techniques and machine learning algorithms. This will support archaeologists in the reconstructing of cultural heritage fragments as they carry out their studies. Some existing works, and the techniques and algorithms applied in them, have been discussed taking note of the impact of these techniques with respect to a high accuracy result, and their limitations. In addition, the potentials of computer vision and machine learning were highlighted by presenting existing works that have applied these techniques, noting the areas where they fall short. This led to the presentation of the potsherd reconstruction case study that this paper is based on. Furthermore, the problem statement of the study was presented, and the approach and method discussed. While the research approach falls under the realist and positivist approach, its application in a social and cultural heritage context is also significant. Hence, the objective of this methodology is to enhance and improve the accuracy of data capture and processing, leading to an accurate and optimal potsherd reconstruction. Finally, the future research directions of this study were presented with the goal of developing a framework that can make image processing a less complicated task within a pottery reconstruction context and, in general, cultural heritage as well.

REFERENCES

- Alpaydm, E. (2010). *Introduction to machine learning* (2nd ed.). The MIT Press.
- Barclay, A., Knight, D., Booth, P., Evans, J., Brown, D. H., & Wood, I. (2016). *A Standard for Pottery Studies in Archaeology*. Medieval Pottery Research Group.
- Colantoni, A., & Gottarelli, A. (2008). A formalized approach to pottery typology: The case of some typical shapes from the Late Bronze Age in Northern Syria. In *5th International Congress on the Archaeology of the Ancient Near East (ICAANE)* (pp. 455–463). Madrid: Academic Press.
- Collett, L. (2012). *An Introduction to Drawing Archaeological Pottery*. Institute for Archaeologists.
- Di Angelo, L., Di Stefano, P., & Pane, C. (2017). Automatic dimensional characterisation of pottery. *Journal of Cultural Heritage*, 26, 118–128. doi:10.1016/j.culher.2017.02.003
- Díaz-Marín, C., Vendrell-Vidal, E., Aura-Castro, E., Abate, A. F., Sánchez-Belenguer, C., & Narducci, F. (2015). Virtual reconstruction and representation of an archaeological terracotta statue. In *2015 Digital Heritage International Congress* (pp. 699–702). Granada: IEEE. 10.1109/DigitalHeritage.2015.7419602
- Drucker, J. (2012). Humanistic Theory and Digital Scholarship. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 85–95). University of Minnesota Press. doi:10.5749/minnesota/9780816677948.003.0011
- Eiteljorg, H. (2004). Computing for Archaeologists. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 20–30). Blackwell., doi:10.1002/9780470999875.ch2
- Forsyth, D. A., & Ponce, J. (2012). *Computer Vision: A Modern Approach* (2nd ed.). Pearson.
- Ghosh, D., & Kaabouch, N. (2016, January). A survey on image mosaicing techniques. *Journal of Visual Communication and Image Representation*, 34, 1–11. doi:10.1016/j.jvcir.2015.10.014
- Haykin, S. (2009). *Neural Networks and Learning Machines* (3rd ed.). Pearson.
- Jeffrey, S. (2010). Resource Discovery and Curation of Complex and Interactive Digital Datasets. In C. Bailey & H. Gardiner (Eds.), *Revisualizing Visual Culture* (pp. 45–60). Ashgate Publishing.
- Kampel, M., & Sablatnig, R. (2003). Virtual reconstruction of broken and unbroken pottery. In *Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM 2003)* (pp. 318–325). Banff: IEEE. 10.1109/IM.2003.1240265
- Makridis, M., & Daras, P. (2012). Automatic Classification of Archaeological Pottery Sherds. *Journal on Computing and Cultural Heritage*, 5(4), 15:1-15:21. doi:10.1145/2399180.2399183
- Marchetti, N., Angelini, I., Artioli, G., Benati, G., Bitelli, G., Curci, A., ... Rocchetti, M. (2017). NEAR-CHOS. Networked Archaeological Open Science: Advances in Archaeology Through Field Analytics and Scientific Community Sharing. *Journal of Archaeological Research*, 1–23. doi:10.1007/10814-017-9112-4
- Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective* (2nd ed.). Taylor & Francis.
- Moeslund, T. B. (2012). Image Acquisition. In *Introduction to Video and Image Processing: Building Real Systems and Applications* (1st ed., pp. 7–24). Springer Science & Business Media., doi:10.1007/978-1-4471-2503-7_2

- Puglisi, G., Stanco, F., Barone, G., & Mazzoleni, P. (2015). Automatic Extraction of Petrographic Features from Pottery of Archaeological Interest. *Journal on Computing and Cultural Heritage*, 8(3), 13:1-13:13. doi:10.1145/2700422
- Rice, P. M. (1987). Special Topics in Archaeological, Ethnoarchaeological and Ethnographic Pottery Studies. In *Pottery Analysis: A Source Book* (pp. 274–306). The University of Chicago Press. doi:10.2307/2804120
- Roman-Rangel, E., Jimenez-Badillo, D., & Marchand-Maillet, S. (2016). Classification and Retrieval of Archaeological Potsherds Using Histograms of Spherical Orientations. *Journal on Computing and Cultural Heritage*, 9(3), 17:1-17:23. doi:10.1145/2948069
- Sabharwal, A. (2015). *Digital Curation in the Digital Humanities*. Elsevier. doi:10.1016/B978-0-08-100143-1.00001-5
- Smith, A. (2004). Preservation. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 576–591). Blackwell. doi:10.1002/9780470999875.ch37
- Smith, P., Bepalov, D., Shokoufandeh, A., & Jeppson, P. (2010). Classification of Archaeological Ceramic Fragments Using Texture and Color Descriptors. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 49–54). San Francisco, CA: IEEE. 10.1109/CVPRW.2010.5543523
- Stamatopoulos, M. I., & Anagnostopoulos, C.-N. (2016). *3D digital reassembling of archaeological ceramic pottery fragments based on their thickness profile*. ArXiv:1601.05824
- Tsiafaki, D., Koutsoudis, A., Arnaoutoglou, F., & Michailidou, N. (2016). Virtual reassembly and completion of a fragmentary drinking vessel. *Virtual Archaeology Review*, 7(15), 67–76. doi:10.4995/var.2016.5910

ADDITIONAL READING

- Daumé, H. (2012). *A Course in Machine Learning*. CIML.
- Hartley, R., & Zisserman, A. (2003). *Multiple View Geometry in Computer Vision* (2nd ed.). Cambridge University Press.
- Oxholm, G., & Nishino, K. (2013). A flexible approach to reassembling thin artifacts of unknown geometry. *Journal of Cultural Heritage*, 14(1), 51–61. doi:10.1016/j.culher.2012.02.017
- Pintus, R., Pal, K., Yang, Y., Weyrich, T., Gobbetti, E., & Rushmeier, H. (2016). A Survey of Geometric Analysis in Cultural Heritage. *Computer Graphics Forum*, 35(1), 4–31. doi:10.1111/cgf.12668
- Rasheed, N. A., & Nordin, M. J. (2015). A Survey of Computer Methods in Reconstruction of 3D Archaeological Pottery Objects. *International Journal of Advanced Research*, 3(3), 712–724.
- Schreibman, S., Siemens, R., & Unsworth, J. (2004) *A Companion to Digital Humanities*. doi: . doi:10.1002/9780470999875
- Shapiro, L., & Stockman, G. (2001). *Computer Vision*. Pearson.

KEY TERMS AND DEFINITIONS

Computer Vision: The ability to make computers or machines to see, interpret, and do similar tasks that humans can do.

Cultural Heritage: An expression of the way of life of a people, inherited and passed on from generation to generation and preserved in the present for posterity. The way of life of a people include its customs, traditions, values, knowledge, belief, law, art, and anything that is acquired and passed on.

Image Acquisition: The process of capturing an unprocessed image from an object or scene by an optical device into a manageable form for processing and analysis purposes.

Machine Learning: The science of getting computers to learn and improve their learning over time in an accurate and automatic manner, by giving them data and information.

Potsherd: A broken piece of pottery, or earthenware generally, with archaeological value.

Pottery: Vessels made from soft clay and hardened by heat.

Pottery Classification: A procedure or process of systematic sorting and grouping together of pottery or potsherd of high similarities and characteristics.

Pottery Documentation: The process of acquiring, classifying, processing, and recording pottery objects, physically or digitally, for preservation and research purposes.

Pottery Reassembly/Reconstruction: The act of restoring potsherds, physically, or digitally, so that it looks as its original condition as much as possible.

ENDNOTE

¹ <https://creativecommons.org/>

Acoustic Presence Detection in a Smart Home Environment

Andrej Zgank

 <https://orcid.org/0000-0002-1411-5293>

University of Maribor, Slovenia

Damjan Vlaj

 <https://orcid.org/0000-0001-7689-0407>

University of Maribor, Slovenia

INTRODUCTION

Rapid research progress in the field of Information Communications Technology (ICT) has an important influence on everyday life, with an increasing number of devices present in the human environment. Existing communication networks and their successors, like 5G, enable the development of a smart home environment, where all these devices communicate and integrate into one system, with the objective to ease and support human life. The population in developed countries is aging fast. This will result in an increased ratio of elderly persons who will need additional care. The smart home environment can provide one of the solutions to prolonging the independent life of the elderly and, in addition, simplify the organization for caregivers from the point when they can't live independently anymore onwards.

A broad set of various services can be implemented in a smart home environment, with the objective to improve the quality of life of the inhabitants. A basic set of services is oriented toward establishing a fully automated and energy-efficient home environment. An extended set of smart home services includes fusion of different information and media content sources into a common platform, which serves as a virtual assistant. The interaction with the assistant can be carried out in an almost natural way if embodied conversational agents (Rojc et al., 2017) are implemented in this role. One of the key pieces of information needed for successful operation should answer the question if users are present in the environment. A smart home environment must be able to monitor the users inside the home, to be aware of their presence and movements in particular rooms.

The presence detection systems in a smart home environment can be implemented in various ways. The basic concept behind presence detection is to measure some of the biological, technical or environmental conditions, which are influenced or generated by the users who are present in the smart home. From the quality of experience point of view, users prefer those presence detection systems which are non-invasive. In the opposite case, users have to carry with them a presence device, which reduces the perceived smart home experience significantly. The presence device can be a dedicated hardware module, which is, typically, using one of the in-door positioning methods to sense the user's presence. Wireless systems like RFID, NFC, Bluetooth, Bluetooth Low Energy (BLE), etc. have been applied successfully for this task. Another approach is to use some terminal equipment that users are carrying with them as a presence device. Nowadays, a smartphone is best suited for this task, although it still has to be carried by the users all the time to fulfill the role of the presence device, which reduces the comfort of smart home inhabitants.

DOI: 10.4018/978-1-7998-3479-3.ch011

The non-invasive methods for presence detection in a smart home can be classified according to the modality which is being measured or monitored to detect the user's presence. A frequently used approach is based on Passive Infra-Red (PIR) motion detection sensors. The PIR sensor is a small electronic device, which operates locally and detects the increased level of infra-red light radiation emitted by humans. Another solution is based on acoustic presence detection, which estimates the users' presence based on uttered words and other sound produced by the speakers during their presence and activities in the environment of a smart home. In this case, the audio signal is captured by a single microphone or by a microphone array, and processed to detect users' presence with audio and speech classification algorithms. A similar approach can be carried out using the visual modality, where a video camera captures a signal and the user's presence is detected with visual object recognition algorithms. An example of the non-invasive presence detection method, which is measuring technical conditions, is based on estimating the change in the electromagnetic field using IEEE 802.11 Wi-Fi access points.

The objective of the proposed encyclopedia article¹ is to present the approach of acoustic presence detection in a smart home environment, analyze the modules which are building the complete system, and propose the key elements for the given context. The acoustic presence detection solution benefits greatly from Human-Computer Interfaces, which are part of intelligent ambient. The usage of acoustic presence detection in a smart home environment is one of the building blocks which is a prerequisite for establishing a natural user-interface for human users.

BACKGROUND AND LITERATURE REVIEW

In the area of presence detection for the smart home environment, passive infra-red sensors are the method which has been applied most frequently for a user's presence detection (Monaci and Pandharipande, 2012). The main reason is its low technical complexity and low costs. The methods based on video modality (Ahvar et al., 2016) were used successfully for presence detection in different smart home set-ups, despite their computational complexity. The third method for sensing a user's presence in the environment is based on acoustic signals. Acoustic presence detection can be carried out with different algorithms. One of the simplest is based on measuring loudness in the smart home environment (Brgulja et al., 2010). More robust and accurate are methods based on low-level features, like the ones extracted by the openSMILE toolkit (Eyben et al., 2013). The most sophisticated acoustic presence detection systems are using classification or recognition algorithms, which are processing acoustic signals captured in a smart home environment by microphone arrays.

Smart environments contain various electronic devices that allow us to interact with human and machine. For detection of the acoustic presence we need a microphone for capturing an acoustic signal in such environments. The installation of the selected microphone in the room is not self-evident, because we need to know the acoustics of the room. The book (Virtanen et al., 2018) provides an overview of the methods for computational analysis of sound scenes and events for different environments, while sound environment analysis in a smart home was provided in (Sehili et al., 2012). The main approach for capturing sound in a smart home environment is microphone array. The study and implementation of microphone arrays are explained in detail in (Brandstein & Ward, 2001). The audio signal captured with microphone arrays is the basis for further steps of acoustic presence detection in smart home environments that can contain different types of acoustic signals, from speech, music, various noises from the environment, and, of course, also a signal of pure silence. If the goal is to detect the presence of a human, then pre-processing techniques must be used to remove the noise from the captured audio signal,

while the music in the background is also intended to be noise. Most of the pre-processing algorithms will work well if a good noise estimate can be provided. This can be achieved by using Voice Activity Detection (VAD) algorithms. The VAD algorithm excludes parts that contain only noise. In this way, a good estimation of the noise characteristics can be obtained, which is then used in the noise reduction procedures. Noise reduction techniques are well presented in the book (Vaseghi, 2009), while a comprehensive presentation and comparison of VAD algorithms is given in (Stadermann, 1999). The VAD algorithm also plays an important role in the classification process, as it removes parts of the captured audio signal that contain noise. In the classification process, therefore, it is not necessary to process the entire captured signal, which makes the classification less computationally demanding and more accurate. Before the classification process, it is also necessary to select the optimal features that will contribute to a good classification result. Most of the feature extraction algorithms are given in (Gupta et al., 2018).

The end result of acoustic presence detection is given by the classification task, which belongs to the area of machine learning and statistical modeling. The classification which is part of acoustic presence detection is frequently based on classification approaches also applied in the area of automatic speech recognition (Junqua and Haton, 2012) and acoustic analysis (Zöhrer and Pernkopf, 2016). Gaussian Mixture Models (GMM) are an example of classification methods which were applied successfully for many scenarios. GMMs were used for classification of speakers' characteristics (Mallouh et al., 2018), such as gender, age, speaking style, etc. The GMM classification approach was also used for language identification (Anjanendu et al., 2018) and music identification (Chaudhary et al., 2019). A related approach (Benzeghiba et al., 2007), used widely for a similar group of tasks are Hidden Markov Models (HMM). HMM are, in the first place, used for speech recognition, but can be also trained for other scenarios (Ma et al., 2006). Deep Neural Networks are currently the state-of-the-art approach for classification of various acoustic categories (Mesaros et al., 2018), and can also be applied successfully for acoustic presence detection in a smart home environment.

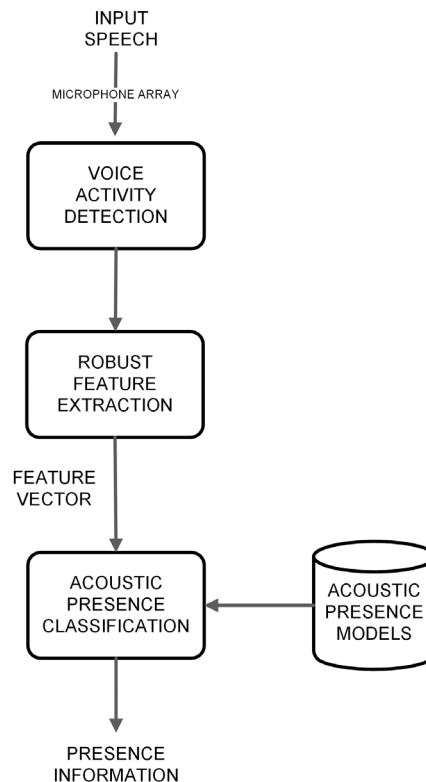
ACOUSTIC PRESENCE DETECTION

The next section will present the functionality and interoperability of acoustic presence detection in the smart home environment from the system point of view (Figure 1). The first step is a good capture of the audio signal, which can be achieved with a good microphone array. However, in the captured audio signal is not present only speech, therefore it is necessary to define the areas where the speech signal is located. For this matter, a voice activity detection algorithm is used. From the captured speech signal, it is necessary to define robust features with the feature extraction algorithm, which will be most appropriate in the classification process for the acoustic presence detection. The classification process takes into account the input feature vectors and the information collected in the acoustic model in a statistical way and then classifies the input signal in the predetermined classes. In the following subsections, each building block will be described in more detail.

In a smart home environment, the presence of people can be detected by capturing a video signal or by capturing an audio signal with different types of microphones. Most people do not feel comfortable in the presence of video cameras, therefore, in smart home environments, the detection of people's presence at the acoustic level of audio signal processing is often preferred.

Acoustic presence detection belongs to the group of non-invasive methods, as no dedicated user's device is needed, which is one of its main advantages. The next important advantage over other methods is that acoustic presence detection also works without direct line-of-sight contact with the user. Thus, it

Figure 1. Acoustic presence detection block scheme

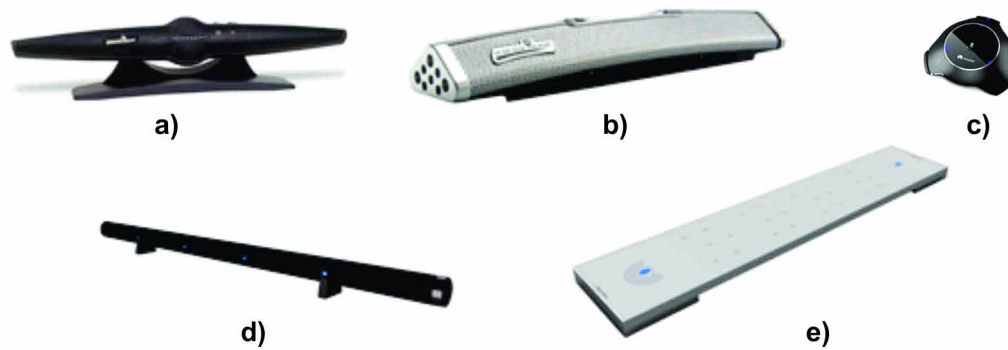


is easier to cover larger areas in the smart home, especially in the case when the rooms are of irregular shape. The acoustic presence detection is also capable of operation in adverse visual conditions (e.g. darkness, low-visibility), where it outperforms systems based on video modality. Another advantage of acoustic presence detection is connected with the system resources needed for operation. Basic acoustic presence detection can work locally, with a limited amount of system resources, which can be provided by an embedded system. The main disadvantage of acoustic presence detection is the fact that it collects the user's spoken interaction and, thus, interferes with the user's privacy. Another disadvantage of acoustic presence detection is connected to various environmental sounds (e.g. background conversation, music streaming, TV, etc.) which are present in the smart home environment and decrease the acoustic presence detection accuracy.

Capturing Signals in a Smart Home Environment

The best results of capturing an audio signal can be achieved by close-talking microphones, but, in smart home environments, this is not useful. In addition, these microphones can be used to develop such systems and to capture reference data. Microphones which do not disturb people are usually placed as a device with a microphone array that captures the audio signal well in the entire room. There are quite a few devices on the market that allow the capture of an audio signal in the room. Figure 2 shows a visual presentation of some of these devices. Below we will do a review of these devices and give an opinion on their use.

Figure 2. Devices with microphone arrays – a) Voice Tracker I from Acoustic Magic, b) Voice Tracker II from Acoustic Magic, c) VPM220 Microphone Array from Huawei, d) Condor Beamforming Array Microphone - MT600 from Phoenix Audio Technologies, and e) Beamforming Microphone Array from ClearOne



The Voice Tracker I (Figure 2.a) (Voice Tracker I, 2018) and the Voice Tracker II (Figure 2.b) (Voice Tracker II, 2018) are produced by the manufacturer Acoustic Magic. They differ in the number of microphones used. The first one uses eight microphones in the length of 45.7 centimeters, while the second one uses six microphones in the length of 25.4 centimeters. The Voice Tracker I has the ability to capture signals up to nine meters, while the Voice Tracker II allows the acquisition of a signal at a shorter distance up to six meters. Both operate in the frequency range between 100 Hz to 11,250 Hz. The VPM220 Microphone Array (Figure 2.c) (VPM220 Series, 2018) is produced by the manufacturer Huawei. This microphone array allows reception of the signal up to a distance of six meters, and its operating frequency range is between 100 Hz and 22 kHz, which is the best of all the devices presented. The Condor Beamforming Array Microphone - MT600 (Figure 2.d) (Condor MT600, 2018) from the manufacturer Phoenix Audio Technologies, uses 15 microphones in the length of 122 centimeters. The range of capturing the signal of this microphone array is nine meters and the operating frequency range is between 50 Hz and 16 kHz. The last of the analyzed microphone arrays is Beamforming Microphone Array (Figure 2.e) (Beamforming Microphone Array, 2018) from manufacturer ClearOne, which uses 24 microphones in the length of 76.2 centimeters. Its range of signal capture is up to three meters, which is the shortest distance between the analyzed microphone arrays, and operates in the frequency range between 150 Hz and 16 kHz. In our opinion, for this microphone array, the lower frequency range is slightly too high, especially if we want to detect the presence of a speaker based on the pitch value of the speaker. All the microphone arrays compared can operate independently, except for the VPM220 Microphone Array (Figure 2.c), which works only in coexistence with the Huawei TE Series HD Videoconferencing Endpoints (TE Series HD, 2018).

Each of the manufacturers suggests the best place to install their device in order to work optimally. Some of the proposed microphone arrays are installed on the conference table, another on the wall, and the third can be placed on the ceiling of the room. The above-mentioned devices have different lengths of reception of the speech signal, so it is not necessary that each device covers the entire room, especially if we have a larger room in mind. The question arises as to how to capture the speech signal well even in the corners of the room itself, and not just near the microphone array. It is useful to detect the presence of the user when he enters the room with the smart environment. Therefore, it would be appropriate for

microphones to be installed at the entrances to the room. The microphones can be placed in devices such as the light switches. In such installation of the microphones room acoustics are especially important. In the next paragraph, we will give guidance on what we need to consider room acoustics successfully, especially when there are several microphones installed in the room.

One of the most important criteria for recording sound in a room is the room's geometry itself. In addition to the shape of the room, the room's volume, the arrangement of furniture and the materials used in the walls, ceiling, floor, and the furniture are also important. Different materials have different absorption characteristics that can suppress the sound level of the audio signal greatly. The contrast of the absorption of the signal is the reflection and also the scattering of the audio signal on various materials, which can, additionally, introduce distortions of the signal source. How to deal with room acoustics is presented in more detail in (Kuttruff, 2017).

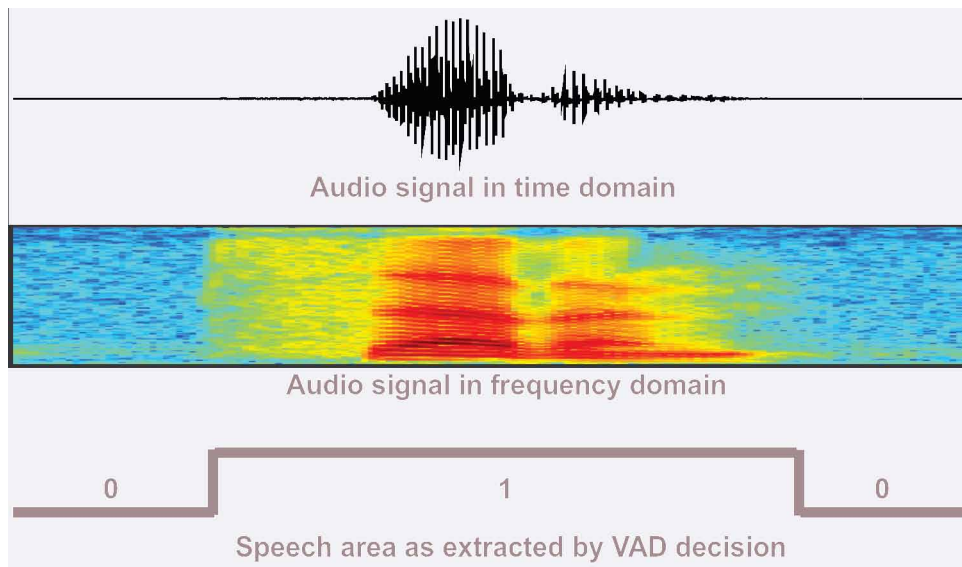
Pre-processing, VAD and Feature Extraction

Once an audio signal is captured, it is necessary to use an effective Voice Activity Detection (VAD) algorithm, which detects the areas of speech in the audio signal. The VAD algorithm can basically distinguish between speech which is mixed with background noise and the background noise itself. In order to achieve the best possible operation of the VAD algorithm, various pre-processing techniques of the captured speech signal must be used. We can use various procedures for speech enhancement and noise reduction, which can improve the quality of the speech signal. These procedures are based on Wiener filtering and spectral subtraction. If, however, we capture a speech signal using a large number of microphones, we can also use techniques such as source separation, channel identification, or speech de-reverberation. All these procedures are described in detail in (Huang et al., 2006).

Existing VAD algorithms try to distinguish between speech and noise based on the information they receive from the frames. In order to use conventional methods of signal processing, it is useful that the signal within a frame is stationary. This can be ensured if the frame duration is between 20 and 30 milliseconds. The result of voice activity detection is the binary value that represents information on whether the input signal contains speech (e.g., the output is 1) or only noise (e.g., the output value is 0). Below, we will make a review of the VAD algorithms from those simple to the more complex ones, which use different features that help to make the right VAD decision.

Most VAD algorithms use features that are based on the calculation of energy or power in the signal. However, since it has been shown that the use of energy as the only feature brings worse results, especially in low signal-to-noise (SNR) environments, in many cases, other features are added. One of these is the zero-crossing rate criterion, which is used in (Kotnik et al., 2001). In combination with the energy of the frame it has improved voice activity detection. The use of the autocorrelation function over the frequency spectrum of each frame also contributed to the improvement of the VAD activity (Vlaj et al., 2004). Some VAD algorithms use as a feature the distance between cepstral features (Haigh & Mason, 1993). In (Renevey & Drygajlo, 2001) spectral entropy was used as a feature. The Gaussian probability distribution model presented in (McAulay & Malpass, 1980), and then used in (Ephraim & Malah, 1984), is the basis of the VAD algorithm based on the statistical model (Vlaj et al., 2012). In (Price et al., 2018), neural networks are used to decide on the presence of speech. In (Zou et al., 2014), the Linear Predictive Cepstral Coefficients (LPCC), that were determined on the basis of autoregressive modeling, were used to decide VAD. In all of the aforementioned VAD algorithms, information was used from the time domain or from the frequency spectrum domain. The presence of the speech area in the audio signal is then given by the VAD decision as shown in Figure 3. In (Eshaghi & Mollaei, 2010), the

Figure 3. Captured audio signal in time and frequency spectrum domain and the area where only the speech signal is present based on the VAD decision



characteristics of the VAD algorithm are derived from the spectrum obtained by the wavelet transformation. VAD algorithms that use higher-order statistics are among the more complex ones (Li et al., 2005).

Most of these VAD algorithms have their advantages as well as weaknesses. Thus, VAD algorithms based on statistical models (Vlaj et al., 2012; Wu et al., 2018) and neural networks (Price et al., 2018) can achieve good results, but they are quite dependent on the training set needed for the model or network to be well trained. Of course, if we have a good training set that is well segmented (the borders between the speech signal and the noise are correctly set), we can get very good results with statistical models and neural networks. For a good segmentation, however, it is usually necessary to annotate the training material of the speech database manually. Recently, attempts have been made to achieve voice activity detection through the help of multimodality (Ariav & Cohen, 2019).

Since the VAD decision is made on the frame bases (for example, the duration of one frame is 25 milliseconds), frequent switching between one and the other decision can occur. Therefore, it is reasonable that in smart home environments we capture the signal at longer time intervals, for example, 250 milliseconds. This solution has proved to be very effective in defining the beginning and ending boundaries of the speech signal in speech-based platforms for intelligent environments (Kos et al., 2018). In the following, we will explain briefly the idea of capturing signals. The speech signal is captured every 250 milliseconds and is saved temporarily in a circular buffer. This 250 millisecond signal is then cut into 20 sub-frames, with 50 percentages overlap. Then, the process of voice activity detection is performed in each of the 20 sub-frames. We can use any of the VAD algorithms that make the decision on the frame basis. If more than 10 sub-frames are detected as speech, the whole frame is recognized as speech; otherwise, the whole frame is recognized as silence or noise. The resulting segmented signal is the first of the indicators for acoustic presence detection in a smart home environment. However, if we want to find out who is present in such an environment, then we have to use the classification methods.

Before the classification procedure it is necessary to create features which are most appropriate for the acoustic presence detection of a speaker in a smart home environment. The speaker can be an adult (male or female) or a child. The feature extraction procedure is a key factor to successful gender or age

classification. Various approaches, such as Mel-Frequency Cepstral Coefficients (MFCCs) (Gaikwad et al, 2012; Yücesoy & Nabiyeu, 2013), prosodic features like pitch (Ramteke et al., 2018), were used for the gender classification task. Also, paralinguistic features, such as speech rate, voice quality, pauses, can be applied successfully to the gender classification task (Mueller, 2006). In other types of classifications, which were not used for gender or age classification but for the classification of human activities in natural environments, features were used such as MPEG-7 Low-Level Descriptors (Ntalampiras et al., 2012) and Perceptual Wavelet Packets (Ntalampiras et al., 2009). After reviewing all the feature extraction algorithms, we would select MFCCs, because in our opinion it would be best appropriate for acoustic presence detection. MFCCs are determined using a discrete cosine transformation from logarithmic values of Mel-filter banks ($\log Fbank$), as determined by the equation below. In this case, K represents the number of MFCCs and J the number of logarithmic values of Mel-filter banks.

$$MFCC(i) = \sum_{j=1}^J \log Fbank(j) \cdot \cos\left(\frac{\pi \cdot i}{J}(j - 0.5)\right) \quad 0 \leq i \leq K \quad (1)$$

Acoustic Classification Procedures

The main objective of the classification task is to distinguish between different acoustic categories present in the signal captured from the smart home environment. In the case of acoustic presence detection, at least two classes must be defined. The first one determines the presence of the human user(s), according to the spoken signal, and the second one determines the presence of silence, which indicates the absence of the user. Acoustic presence detection systems are deployed frequently in an intelligent ambient, where other sources of audio also exist. One such example of audio sources is the user's media devices, like TV or home entertainment/music system. An audio signal produced by them can be misclassified as the user's signal, and, as a consequence, misinterpreted as a user's presence by the smart home main control and the decision module. One of the possible solutions to reduce the proportion of such misclassification is to introduce an additional separate class, which determines the various audio background effects which can frequently be found in the user's environment. In this case, it is beneficial if the type of user's environment can be specified precisely, as the categories of acoustic background events and sources differ highly between them.

Acoustic presence detection can be applied in conjunction with automatic speech recognition, as the key elements share a large portion of algorithms. We will, thus, first present classification approaches which are also used frequently in the domain of human language technologies. Gaussian Mixture Models (GMM) are a statistically based approach. Each acoustic class has its own separate GMM model, which is defined as:

$$p(x) = \sum_{i=1}^K \phi_i \mathcal{N}(x | \mu_i, \sigma_i) \quad (2)$$

The GMM model is defined by a mixture of Gaussian probability density functions, where weight ϕ_i specifies the impact of each particular probability density function. The function is defined by its mean value μ_i and variance σ_i . The total number of mixtures used in a GMM depends on two factors. The first one is the amount of acoustic training material belonging to a particular audio class. The second

one is the distribution and sparseness of extracted feature vector values in the multi-dimensional feature space. Typical GMM models can have several 10 to several 100 probability density functions. During the classification procedure, each acoustic frame with feature vectors gets classified, to estimate, which acoustic class it belongs to.

The output of the classification approach is a stream of classes' labels, where short glitches are also present. In the case of a smart home environment, such short glitches usually present misclassified input, as it is highly unlikely that the user leaves the monitored area for as short a duration as 10 or 20 ms (a glitch with a duration of 1 or 2 input windows). To minimize this effect, a post-processing procedure can be applied to the output stream of labels, which is used to smooth out these short glitches.

Another classification approach, which has its origins in the field of automatic speech recognition, is based on Hidden Markov Models (HMM). The HMM approach also belongs to the category of statistical modeling. A left-right models' topology with continuous Gaussian probability density functions is being used for acoustic modeling. The number of states per model varies between 2 and approx. 10, and it depends on the characteristics of the acoustic data used for training. The HMM acoustic models are defined as:

$$\lambda = (A, B, \pi) \quad (3)$$

Where λ denotes the HMM model, A denotes the transition probability between state i and the next state j , B denotes an observation probability matrix, and π denotes the initial state distribution. Each HMM state is modeled by a continuous density probability function. In the case when a Gaussian function is applied, the weight ϕ_i , mean value μ_i , and variance σ_i define the function. The Baum-Welch re-estimation procedure is applied for training the acoustic models. The trained acoustic models are used for classification with the Viterbi decoder, which estimates the maximum probability that the input feature vector results in a particular acoustic environment class. The estimation takes place on larger parts of the input signal, which are homogenous from the voice activity detection point of view. As a consequence, the short glitches, which are present in the case of GMM classification, are absent in the case of HMM modeling, which presents an advantage over the method presented above. An additional advantage of HMM modeling is that it can share modules with automatic speech recognition if this uses the same architectural approach. The disadvantage over the GMM method is its computational complexity, which is higher than for the GMM approach.

With the availability of larger acoustic data sets and multi-core Graphical Processing Units (GPU), various deep learning methods could also be applied for the acoustic presence detection task. One of the most commonly used ones is Deep Neural Networks (DNN). In this case, each acoustic class is needed for presence detection, represented by a neural network. Recently (Xia et al., 2019), the convolutional neural networks (Valenti et al., 2016), which originate from the area of image processing, present the most advanced approach for acoustic classification. In this case the log-mel spectrograms are used to present the input audio signal. The ensemble method (Xu et al., 2018) provide a way, how different classification approaches can be successfully combined in a complex solution with increased accuracy.

A classification approach which has its origins in other areas of machine learning is Support Vector Machines (SVM), which belongs to the category of supervised learning models. The extracted feature vectors are, in the case of SVM, grouped in classes in a way to build groups separated by a hyperplane in a multidimensional feature space. The objective is to maximize the distance between the hyperplane and the nearest feature vector element.

The advantage of acoustic presence detection is also its ability to use the input speech signal to extract other important information, which is given in the uttered speech implicitly or explicitly. In the case when, besides the acoustic presence, the system also detects other users' characteristics, more detailed separation is needed of acoustic training material. An example of implicit information given in the uttered speech which can be then classified is various speakers' characteristics. The more interesting ones belong to a demographic category, such as the speaker's gender and/or age. This information can be valuable in a smart home environment, as it could give an additional value to assisting services and human-computer interfaces.

The inclusion of more information in the acoustic presence detection system requires increased complexity of classifiers. The most straightforward approach is to increase the number of modeled classes according to the classification task. Another approach is based on system fusion, where different classifiers are combined to provide one final decision regarding the output result. The explicit information given in the uttered speech signal is the textual information. This one can be processed with an automatic speech recognition system, which is part of the advanced human-computer interface in a smart home environment. As the end result, the automatic speech recognition output also gives valuable information about the speaker's presence. The decision whether to use automatic speech recognition as a form of acoustic presence detection, should bear in mind the user's privacy concerns, which are still important, despite the wide usage of different social networks and new methods of communication.

SMART HOME SERVICES BASED ON ACOUSTIC PRESENCE DETECTION

Services represent the key results of the smart home environment. To be able to improve the services' operation, advanced functionalities are required. One such example is the customization of available media content according to the detected user's age, which can be used as a type of parental content control for children. The speakers' demographic characteristics can be also used in other types of intelligent ambient modules, which have to deal with the media or information. The media content recommendation system, which is frequently included into a video on demand service or music/radio streaming service, can benefit from information about the user's age and gender information, each time the system must be initialized or cold-started. A further example is when the recommendation system must operate without the explicit knowledge of the user's identity, or in cases when it operates in a public or private environment where the users' population changes frequently. Virtual assistants are one of the possible human-computer interface services in the smart home, which provide a way of natural communication to the users. The user's characteristics can be important additional information for a virtual assistant, especially in those cases where the virtual assistant uses only the tactile and/or haptic input modality. In the case when an embodied conversational agent gets applied in the role of the virtual assistant, this information can be used for the initial part of the communication, or, in cases when the degraded acoustic environment reduces the accuracy of direct audio input signal to the embodied conversational agent.

Another example of advanced services' functionality is the personalization of a graphical user interface in a way to use larger icons and font size in the case when the acoustic presence detection system classifies the input audio signal as one belonging to an elderly person. A more sophisticated version of an acoustic presence detection system can run a speaker detection system, which could detect the particular user in the monitored environment. The drawback of this solution is its complexity, and the fact that, for a typical home scenario, the method produces too detailed information, as the same personalization/customization result can be achieved using simpler methods and characteristics, even without collect-

ing sensitive personal information. The typical smart home users prefer the approaches where their personal information, including voice signals, are processed locally, without the need to be uploaded to the cloud-based infrastructure for further processing. An additional benefit is also the possibility to develop the architecture as a stand-alone system, with the possibility to carry out the operation, even in the case when the internet connection suffers from black-out.

DISCUSSION AND FUTURE TRENDS

After analyzing the characteristics of the compared audio capture devices, the Condor Beamforming Array Microphone - MT600 from manufacturer Phoenix Audio Technologies selection would be proposed to capture the audio signal up to nine meters, and has a wide frequency range between 50 Hz and 16 kHz. This is the only microphone array that has a lower frequency limit set to 50 Hz. Such low limit also makes it possible to determine the pitch value of men with a deeper voice. The microphones, that should be placed in the switches, must be able to capture an audio signal in a similar frequency range. We also propose the use of the VAD algorithm, which would, in smart home environments, allow decisions on the presence of speech in the captured signal over extended time, such as 250 milliseconds. For the classification procedures, according to the experience of the previous research, we would propose the use of features that would be determined from the combination of Mel-Frequency Cepstral Coefficients (MFCC) and MPEG-7 Low-Level Descriptors.

The future trends in the area of classification will be oriented toward providing advanced classification algorithms for acoustic presence detections which are capable of operating in a distributed way on low-resource devices, such as embedded IoT systems in wireless acoustic sensor networks. This aspect of smart home environment processing is particularly important from the point of view of taking care for the user's privacy and, consequently, establishing a processing system where all personal information gets processed locally.

CONCLUSION

The article presented details methods of how the user's presence in a smart home environment can be detected from the acoustic signals produced by humans. First, the solutions were analyzed and presented to capture high-quality acoustic signals. The focus was given to different microphone arrays, which clearly outperform single-channel stand-alone microphone solutions. The next important step is oriented toward improving the robustness, and to extract features present in the captured acoustic signal. The last part presents the classification of users' presences, or, in the case of advanced systems, to determine additional users' characteristics.

The acoustic presence detection based on advanced classification approaches can provide information on users' presence successfully to the main smart home module, where other services then benefit from this additional knowledge. This is important for improving the quality of experience in the smart home environment, as services' personalization and customization can be carried out.

REFERENCES

- Ahvar, E., Daneshgar-Moghaddam, N., Ortiz, A. M., Lee, G. M., & Crespi, N. (2016). On analyzing user location discovery methods in smart homes: A taxonomy and survey. *Journal of Network and Computer Applications*, 76, 75–86. doi:10.1016/j.jnca.2016.09.012
- Anjanendu, C., George, A., & Mary, L. (2018). Language Identification Using Gender Dependent GMM-UBM for Three Indian Languages. In *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 510-513). IEEE. 10.1109/ICOEI.2018.8553783
- Ariav, I., & Cohen, I. (2019). An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(2), 265–274. doi:10.1109/JSTSP.2019.2901195
- Beamforming Microphone Array. (2018, December 20). http://www.clearone.com/products_beamforming_mic_array/
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouviet, D., ... Rose, R. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11), 763–786. doi:10.1016/j.specom.2007.02.006
- Brandstein, M., & Ward, D. (Eds.). (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-662-04619-7
- Brgulja, N., Kusber, R., & David, K. (2010). Validating Context Information in Context Aware Systems. In *Intelligent Environments (IE), 2010 Sixth International Conference on* (pp. 140-145). IEEE. 10.1109/IE.2010.33
- Chaudhary, S. R., & Kakarwal, S. N. (2019). Various Approaches in Musical Instrument Identification: A Review. *International Journal of Applied Evolutionary Computation*, 10(2), 1–7. doi:10.4018/IJAEC.2019040101
- Condor MT600. (2018, December 20). *Microphone array*. <https://www.phnxaudio.com/condor/mt600/>
- Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(12), 1109 – 1121.
- Eshaghi, M., & Mollaei, K. (2010). Voice activity detection based on using wavelet packet. *Digital Signal Processing*, 20(4), 1102–1115. doi:10.1016/j.dsp.2009.11.008
- Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 835-838). ACM. 10.1145/2502081.2502224
- Gaikwad, S., Gawali, B., & Mehrotra, S. C. (2012). Gender identification using SVM with combination of MFCC. *Advances in Computational Research*, 4(1), 69–73.
- Gupta, D., Bansal, P., & Choudhary, K. (2018). The State of the Art of Feature Extraction Techniques in Speech Recognition. In S. Agrawal, A. Devi, R. Wason, & P. Bansal (Eds.), *Speech and Language Processing for Human-Machine Communications* (Vol. 664). Springer. doi:10.1007/978-981-10-6626-9_22

Haigh, J., & Mason, J. S. (1993). A voice activity detector based on cepstral analysis. *Eurospeech'93 Proceedings*, 1103 – 1106.

Huang, Y. A., Benesty, J., & Chen, J. (2006). *Acoustic MIMO Signal Processing. Part of the Signals and Communication Technology book series*. Springer. doi:10.1007/978-3-540-37631-6

Junqua, J. C., & Haton, J. P. (2012). *Robustness in automatic speech recognition: fundamentals and applications* (Vol. 341). Springer Science & Business Media.

Kos, M., Rojc, M., Žgank, A., Kačič, Z., & Vlaj, D. (2018). A speech-based distributed architecture platform for an intelligent ambience. *Computers & Electrical Engineering*, 38, 818–832. doi:10.1016/j.compeleceng.2017.07.010

Kotnik, B., Kačič, Z., & Horvat, B. (2001). A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. *Eurospeech'01 Proceedings*, 197 – 200.

Kuttruff, H. (2017). *Room Acoustics* (6th ed.). CRC Press.

Li, K., Swamy, M. N. S., & Ahmad, M. O. (2005). An improved voice activity detection using higher order statistics. *IEEE Transactions on Speech and Audio Processing*, 13(5), 965–974. doi:10.1109/TSA.2005.851955

Ma, L., Milner, B., & Smith, D. (2006). Acoustic environment classification. *ACM Transactions on Speech and Language Processing*, 3(2), 1–22. doi:10.1145/1149290.1149292

Mallouh, A. A., Qawaqneh, Z., & Barkana, B. D. (2018). New transformed features generated by deep bottleneck extractor and a GMM-UBM classifier for speaker age and gender classification. *Neural Computing & Applications*, 30(8), 2581–2593. doi:10.1007/00521-017-2848-4 PMID:30363735

McAuly, R. J., & Malpass, M. L. (1980). Speech enhancement using soft decision noise suppression filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 137 – 145.

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., & Plumbley, M. D. (2018). Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(2), 379–393.

Monaci, G., & Pandharipande, A. (2012). Indoor user zoning and tracking in passive infrared sensing systems. In *EUSIPCO 2012* (pp. 1089-1093). IEEE.

Mueller, C. (2006). Automatic recognition of speakers' age and gender on the basis of empirical studies. *Proceedings of INTERSPEECH'06 – ICSLP*, 2118–2121.

Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2009). Exploiting Temporal Feature Integration for Generalized Sound Recognition. *EURASIP Journal on Advances in Signal Processing*, 2009(1), 807162. doi:10.1155/2009/807162

Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2012). Acoustic Detection of Human Activities in Natural Environments. *Journal of the Audio Engineering Society*, 60(9), 686–695.

Price, M., Glass, J., & Chandrakasan, A. P. (2018). A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks. *IEEE Journal of Solid-State Circuits*, 53(1), 66–75. doi:10.1109/JSSC.2017.2752838

Ramteke, P. B., Dixit, A. A., Supanekar, S., Dharwadkar, N. V., & Koolagudi, S. G. (2018). Gender Identification From Children's Speech. *Proceedings of Eleventh International Conference on Contemporary Computing (IC3)*.

Renevey, P., & Drygajlo, A. (2001). Entropy Based Voice Activity Detection in Very Noisy Conditions. *Eurospeech'01 Proceedings*, 1887 – 1890.

Rojc, M., Mlakar, I., & Kačič, Z. (2017). The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm. *Engineering Applications of Artificial Intelligence*, 57, 80–104. doi:10.1016/j.engappai.2016.10.006

Sehili, M. A., Lecouteux, B., Vacher, M., Portet, F., Istrate, D., Dorizzi, B., & Boudy, J. (2012). Sound Environment Analysis in Smart Home. In F. Paternò, B. de Ruyter, P. Markopoulos, C. Santoro, E. van Loenen, & K. Luyten (Eds.), *Lecture Notes in Computer Science: Vol. 7683. Ambient Intelligence*. Springer.

Stadermann, J. (1999). *Sprach/Pause-Detektion in der automatischen Spracherkennung. Diplomarbeit*. Philips Forschungslaboratorien.

TE Series HD. (2018, December 20). *Videoconferencing Endpoints*. <https://e.huawei.com/in/products/cloud-communications/telepresence-video-conferencing/endpoints/te-hd/>

Valenti, M., Diment, A., Parascandolo, G., Squartini, S., & Virtanen, T. (2016). DCASE 2016 acoustic scene classification using convolutional neural networks. In *Proc. Workshop Detection Classif. Acoust. Scenes Events* (pp. 95-99). Academic Press.

Vaseghi, S. V. (2009). *Advanced Digital Signal Processing and Noise Reduction* (4th ed.). John Wiley & Sons Ltd.

Virtanen, T., Plumbley, M. D., & Ellis, D. (Eds.). (2018). *Computational Analysis of Sound Scenes and Events*. Springer International Publishing. doi:10.1007/978-3-319-63450-0

Vlaj, D., Kačič, Z., & Horvat, B. (2004). Voice Activity Detection Based on Autocorrelation of the Frequency Spectrum. *Elektrotehniški vestnik*, 71(4), 165 – 170.

Vlaj, D., Kačič, Z., & Kos, M. (2012). Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria. *Computers & Electrical Engineering*, 38(6), 1820–1836. doi:10.1016/j.compeleceng.2012.09.003

Voice Tracker, I. (2018, December 20). *Product Details*. <https://www.acousticmagic.com/products/voice-tracker-i-details/>

Voice Tracker, I. I. (2018, December 20). *Product Details*. <https://www.acousticmagic.com/products/voice-tracker-ii-details/>

VPM220 Series. (2018, December 20). *Omnidirectional Intelligent Microphone Arrays*. <https://e.huawei.com/en/products/cloud-communications/telepresence-video-conferencing/peripherals/vpm220/>

- Wu, X., Zhu, M., Wu, R., & Zhu, X. (2018). A Self-adapting GMM based Voice Activity Detection. *Proceedings of 23rd International Conference on Digital Signal Processing*. 10.1109/ICDSP.2018.8631856
- Xia, X., Togneri, R., Sohel, F., Zhao, Y., & Huang, D. (2019). A Survey: Neural Network-Based Deep Learning for Acoustic Event Detection. *Circuits, Systems, and Signal Processing*, 1–21.
- Xu, Y., Kong, Q., Wang, W., & Plumbley, M. D. (2018). Large-scale weakly supervised audio classification using gated convolutional neural network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 121-125). IEEE. 10.1109/ICASSP.2018.8461975
- Yücesoy, E., & Nابیyev, V. V. (2013). Gender identification of a speaker using MFCC and GMM. *Proceedings of ELECO*, 13, 626–629.
- Zöhrer, M., & Pernkopf, F. (2016). Gated recurrent networks applied to acoustic scene classification and acoustic event detection. *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*.
- Zou, Y. X., Zheng, W. Q., Shi, W., & Liu, H. (2014). Improved Voice Activity Detection based on support vector machine with high separable speech feature vectors. *Proceedings of 19th International Conference on Digital Signal*, 763 – 767. 10.1109/ICDSP.2014.6900767

ADDITIONAL READING

- Benesty, J., Chen, J., & Huang, Y. (2008). *Microphone Array Signal Processing*. Springer-Verlag Berlin Heidelberg.
- Caivano, D., Fogli, D., Lanzilotti, R., Piccinno, A., & Cassano, F. (2018). Supporting end users to control their smart home: Design implications from a literature review and an empirical investigation. *Journal of Systems and Software*, 144, 295–313. doi:10.1016/j.jss.2018.06.035
- Everest, F. A., & Pohlmann, K. (2009). *Master Handbook of Acoustics* (5th ed.). McGraw-Hill Companies, Inc.
- Gold, B., Morgan, N., & Ellis, D. (2011). Feature Extraction for ASR. In *Book Speech and Audio Signal Processing*. Processing and Perception of Speech and Music. Wiley Telecom. doi:10.1002/9781118142882.ch22
- Sarma, B. D., & Prasanna, S. M. (2018). Acoustic–phonetic analysis for speech recognition: A review. *IETE Technical Review*, 35(3), 305–327. doi:10.1080/02564602.2017.1293570
- Seltzer, M. L., Yu, D., & Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *ICASSP 2013* (pp. 7398-7402). IEEE. doi:10.1109/ICASSP.2013.6639100
- Virtanen, T., Singh, R., & Raj, B. (2012). Voice Activity Detection, Noise Estimation, and Adaptive Filters for Acoustic Signal Enhancement. In *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley Telecom.
- Zaidan, A. A., & Zaidan, B. B. (2018). A review on intelligent process for smart home applications based on IoT: Coherent taxonomy, motivation, open challenges, and recommendations. *Artificial Intelligence Review*, 1–25.

KEY TERMS AND DEFINITIONS

Classification: An approach to separate data into classes.

Feature Extraction: An algorithm to extract distinctive values from signal.

Microphone Array: Acoustic capturing device, which combines multiple microphones.

PIR Sensor: Presence detection sensor, based on infra-red light.

Room Acoustic: Acoustic characteristics of an environment.

Smart Home: An advanced living environment, where ICT solutions support inhabitants.

VAD: An approach of detecting speech in signal.

ENDNOTE

- ¹ This work is partially funded by the European Regional Development Fund and Republic of Slovenia: project IQHOME.

Section 2

Computational Science

Sense and Boundaries of Computer Simulations

2

Georgios O. Papadopoulos

National and Kapodistrian University of Athens, Greece

Apostolos Syropoulos

 <https://orcid.org/0000-0002-9625-1482>

Greek Molecular Computing Group, Greece

INTRODUCTION

It is generally well established that a good many phenomena of everyday life are quite complicated to be described in a simple, concise and quantitative way. Many examples can spring to mind; from the motion of an object in the Earth's atmosphere to the (long term) weather forecasting, or from bacteria growing experiments in a Petri plate to space craft missions, etc. Indeed, almost all the physical (real) systems, say from natural sciences (e.g., physics, chemistry, biology, geology etc.) to applied sciences (e.g., engineering, economics etc.), and from health sciences to humanities are susceptible to a – more or less – multi parametric and complex prototype.

The aforementioned complexity is due to the following reason: The magnitude of the system under consideration is beyond any reasonable reach and completely unlikely to be reproduced in the laboratory for study

- a) either at a local level; indeed, for example, although it is quite simple to deal with the motion of a gas molecule, it is extremely difficult for a human to do the same with the motion/behaviour of an enormous amount of molecules interacting with each other;
- b) or at a global level, for example, the weather forecast in meteorology, the galaxy collisions in Relativistic Astrophysics, etc., for obvious reasons.

Conceptually, case (a) is by default attacked by implementing statistical methods, while case (b) is attacked by multi parametric analytical systems. Nevertheless, both cases are dealt with the use of computers. In other words, the previously mentioned multi parametric, complex prototype is realised as a (imperative) computer code of a functional nature: some input is given and an output is expected (thus the imperative character); also same inputs produce same results (hence the functional character). This computer code incarnates the term “*computer simulation*” [see (Osais, 2017) for an overview of the programming aspects of simulation].

Definition 1. A computer simulation is a computer program that represents the dynamic responses of one system by the behaviour of another system modelled after it.

BACKGROUND

Computers can be used to simulate physical systems but is the use of computers (and thus the need for computer simulations) something unavoidable? The answer to this question is affirmative.

Indeed for the case (a), general theorems from Statistics dictate that the larger the magnitude of the prototype the more accurate and with less deviation is the forecast for the behaviour of the system. Within this sector of this type of problems the need for the use of random numbers (expressing an arbitrary possible state of the system) is a *sine qua non*. Of course, it is practically impossible for a human to produce massively random numbers and thus to emulate a plausible (physical) state of the statistical system under consideration. At this point the computers enter the study; they can produce in (almost/relatively) no time huge quantities of arbitrary numbers, therefore they suffice for the description of the collection of physical states. Yet, there are two limitations here:

- On one hand, computers produce pseudo random numbers and not completely random ones. This has to do with the whole concept of what a human made machine is capable of. Modern techniques have been tried to smooth away this problem by implementing characteristics of the machine (like the temperature of the CPU or the “entropy” of the accumulated information etc.) which could be used to produce truly random numbers, but the very essence of the problem remains since in any case correlating functions are involved.
- On the other hand, computers do not cope well with real numbers, but only with integers. Thus, two real numbers which are supposed to be equal, do not have their difference equal to zero but rather equal to a given (very) small number. For instance, 0.99999 and 1 can be consider as equal (if the tolerance/accuracy is, say, 0.00001) or not –depending both on the hardware architecture and the software constraints.

Now, for the case (b), the arguments go as follows. There exist systems:

- § which either it is beyond any hope (for various objective reasons) for them to be reproduced in the laboratory and thus to study all the aspects/parameters of the problem, for example, the collision of two black holes in Relativistic Astrophysics or
- § which can be considered as showing extreme behaviour that belongs to the limits of a theory. In this case, it is possible for those limits of the theory/prototype under consideration to surpass the current technical abilities. Then, a computer based prototype can fill in the gap between current technology and theoretical demands/inquiries.

Again, as before, the aforementioned limitations are still valid plus one more; the accuracy of the computer based prototype –leaving aside the issue of potential human errors.

Thus, the general conclusion is that for various different reasons, depending of the case under consideration, the implementation of computer simulations is of fundamental importance.

Let us present a naive example that can serve as a comprehensive illustration of the need for computer simulations.

The Problem: *the study of the electromagnetic radiation near the accretion disks in the vicinity of a black hole* [see (Thorne, 1994) for an accessible account of black holes and (Romero & Vila, 2014) for a recent account of the physics of black holes].

The Solution: (towards the computer simulation)

First, one chooses the underlying theory which will describe the physical phenomenon. This theory is going to be the General Relativity and the space will be parameterized by the Kerr space time geometry. The needed equations (i.e., the actual prototype) are the geodesic equations (i.e., equations describing the generalised “straight” lines in a curved space) – a set of entangled, non linear, ordinary differential equations of the second order on the Kerr background. Thus the conceptually fundamental quantity here is the notion of the derivative (say, with respect the “time” variable).

Second, one has to consider how to realise the notion of the time derivative in a manner proper for a routine treatment. The basic idea is to consider space as a discrete net endowed with a tessellation of given (say) cubes. The edges of the cubes correspond to the three spatial coordinates while the position of a cube on the net determines the time coordinate. The discrete character is very important for two successive points on the net will correspond to a difference of two given cubes, that is, the derivative of space coordinates with respect to time. This procedure is called the discretization *of the continuum* (in general, discretization is the process of transferring continuous functions, models, variables, and equations into discrete counterparts).

Next, the fundamental equations (i.e., the geodesic equations) assume a much more ugly –yet completely “comprehensive” by the computer– form; they become sequences of discrete differences on the net. In fact this very nature of those sequences guaranties the imperative and the functional character of the computer simulation. Now one is ready to provide some initial and/or boundary conditions in the form of data concerning the spatial coordinates and their first time derivatives (i.e., velocities) at specific areas of the net. Then, what one has to do is just to observe the (“time”) evolution of the system in the form of the sequence terms on the net.

In what follows we are going to present a number of simulation technologies that are useful per se or can be used to solve real life problems. In addition, we are going to discuss the necessity of quantum computing and the limits of simulation. We will conclude with a brief description of the Simulation Hypothesis.

ARTIFICIAL LIFE

Artificial life is a form of computer simulation that is often used to study essential properties of living systems that include evolution and adaptive behaviour [see (Adami, 1998) for a rather complete overview of the field]. Artificial life is a field that owes its name to the American computer scientist Christopher Langton. This field covers a wide range of disciplines, from engineering over biochemistry to physics, biology, and computer science. The ultimate goal of artificial life is to understand the general principles that govern living beings, and to construct life in an artificial medium so that its properties can be compared to life with which we are familiar. Naturally, as a field that is essentially a form of simulation, artificial life aims to show that modern technology can help us to understand life as a phenomenon. Since each scientific field require some sort of foundations, artificial life is no exception and its roots are in Turing’s Automata Theory and the logical, self-reproducing automata of von Neumann. John Horton Conway created a fantastic solitaire game he called “Life” [see (Gardner, 1983) for an accessible description of the game].

A cellular automaton is a feedback machine that consists of an array of cells. These arrays are either one-dimensional, in which case cells are lined up like a chain, or two-dimensional, in which case cells form a grid. The Game of Life is actually a simulation game because it resembles the evolution of a society of living organism. The evolution rules of the Game of Life are based on the following principles:

- i. There should be no way to tell beforehand whether the population can grow without limit.
- ii. It should be possible to set up a system that apparently grows with no limit.
- iii. After some initial growth a population should fade away, settle into a stable state and remain unchanged thereafter, or it should enter an oscillating phase in which it repeats an infinite cycle of two or more periods.

These rules make sure that the behaviour of the population will be interesting and unpredictable. The evolution rules follow:

- i. Every cell that has two or three alive neighbours survives for the next generation.
- ii. A cell that has four or more alive neighbours dies. Also, if a cell has only one neighbour or no neighbours dies from isolation.
- iii. A dead cell that has three or more alive neighbours is resurrected.

GENETIC ALGORITHMS

One natural phenomenon that is most frequently tackled today is evolution: What are the principles by which life produces new, more complicated and more fit life forms that replace existing ones? Evolution is something that is of paramount interest to biologists but engineers are also interested to it mainly because they would like to realize evolutionary technologies that could create better tools and machines from existing ones.

Genetic algorithms are common computational model that (tries to) mimic evolution in order to solve a number of problems. In general, a computational solution of a problem can be perceived as a search through a space of potential solutions. Clearly, we are interested in the “best” solution, thus a computational solution can be seen as an optimization process. When the solution space is large, one has to employ “clever” techniques. One such technique is the genetic algorithms. These algorithms are stochastic processes that model genetic inheritance and the clash of species and individuals for survival [see (Michalewicz, 1996) for a classical exposition of genetic algorithms]. An *evolution program* implements a genetic algorithm. An evolution program maintains a population of individuals,

$$P(t) = \{x'_1, \dots, x'_n\},$$

for each iteration t . Each individual represents a solution to the problem and it is evaluated to see how fit it is. Then, a new population is generated by selecting the more fit individuals and by modifying certain individuals using some sort of “genetic” transformation. Below we give the general structure of an evolution program as it appears in (Michalewicz, 1996):

```
void evolution_program(void) {  
    t = 0;  
    initialize(P(t));  
    evaluate(P(t));  
    while (! terminal_condition) {  
        t++;
```

```

select P(t) from P(t-1);
alter(P(t));
evaluate(P(t));
}
}

```

SWARM INTELLIGENCE

One could say that *swarm intelligence* (Kennedy & Eberhart, 2001) is about intelligence being social. An extreme example of this social intelligence is the brain of a typical human. The brain is composed by approximately 86,000,000,000 neurons (i.e., brain cells) and although no single cell is intelligent, the totality of neurons, which make up our brain, is obviously intelligent. However, swarm intelligence is about bee hives and ant colonies. One could argue that emergence, which is roughly the idea that some properties of a system are irreducible, plays a central role in swarm intelligence. Thus an ant or a bee has some intelligence but the insect alone cannot achieve what a particular totality can achieve.

In principle, *difficult problems* cannot be solved so to yield answers quickly, still there are approximate methods that can yield results quite quickly. These methods are not general solutions to problems and are known as *heuristics*. A *metaheuristic* is a set of ideas and concepts that can be used to define heuristic methods applicable to a wide set of different problems. The Ant Colony Optimization (ACO) (Dorigo & Stützle, 2004) metaheuristic is based on the idea that a colony of artificial ants cooperate in order to find good solutions to some sort of difficult problems.

VIRTUAL REALITY

Definition 2. Virtual reality (VR), is a computer technology that replicates an environment, real or imagined, and simulates a user's physical presence and environment in a way that allows the user to interact with it. It is an emerging trend which is now bringing more public attention. (Joseph, 2016)

Clearly, VR can be used to model and to simulate real environments. In addition, one or more persons can interact with these environments using interactive devices that can send and receive information. Typical examples of such devices are headsets, gloves, or body suits. Also, the user is wearing a helmet that allows her to have stereoscopic view of the simulated environment. Thus one feels as if she is inside this virtual environment. For example, a user can climb stairs, open doors, etc. Special data gloves equipped with force-feedback devices allow users to have the sensation of touch.

LIMITS OF COMPUTER SIMULATION

From the discussion so far one may get the impression that anything is possible with computers. One can even go farther and claim that even the whole universe can be simulated by a computer (Hayes, 2003). Roughly, Hayes investigated the possibility to construct a computer program written in some ordinary computer programming language (e.g., Java or C++) capable of simulating the Universe (whatever one perceives as such). He has reached the conclusion that although this is a particularly difficult task,

nevertheless, it is a feasible one. Now, this conclusion has a number of interesting consequences: Let us assume that we have at our disposal a computer capable of running such a computer program. Then, after some (long?) time, one will witness the emergence of intelligent life forms in this virtual universe, which, quite possibly, on one (virtual) day will construct a (virtual) computer program and a (virtual) computer capable of simulating their own virtual universe, etc. Such a story is an ideal for scripts for sci-fi movies (e.g., the movie “The Thirteenth Floor,” see <https://www.imdb.com/title/tt0139809/> for movie details, is based in a very similar idea), but it has nothing to do with reality [see (Syropoulos, [0910.2859] Can we debug the Universe?, 2009) for detailed discussion.]

Being able to study physical systems *in silico* is something that depends on what the real capabilities of (ordinary) computers are. Recently, Ringel and Kovrizhim (Ringel & Kovrizhin, 2017) demonstrated (but not *proved*) that there are quantum systems that cannot be effectively simulated by computers, even quantum computers! Of course, the first part of this conclusion is no news. Pour-El and Richards (Pour-El & Richards, 1989) have shown that there is non-computability in ordinary physical systems, such as wave-propagation. Also, the first person who demonstrated that (ordinary digital) computers cannot be used to effectively simulate quantum systems is the famous American theoretical physicist and Nobel prize laureate Richard Phillips Feynman. Around 1982, he published a paper (Feynman, 1982) where he pondered about the possibility of simulating a quantum system with a computer. Feynman examined whether it is possible to simulate and not to imitate time. He concluded that time cannot be simulated but it can be imitated because the “simulation” involves a transition from one state to another. Next, he examined whether it is possible to simulate probabilities. Feynman concluded that probabilities can be simulated by two different methods—the computation of probabilities by an ordinary system or the use of a probabilistic computer. He concluded that the first approach does not work because of intrinsic limitations of computer hardware. A probabilistic computer can be used to yield some statistical result after running the computer a number of times. The problem is that nature is unpredictable and when one has a large number of parameters it makes no or little sense to employ a probabilistic computer in the hope that on a run it will predict the real probabilities. Feynman concluded that only a *quantum simulator* should be able to simulate Nature. Quantum simulators are now known as quantum computers. In the light of results like those of Ringel and Kovrizhim (Ringel & Kovrizhin, 2017), one can safely conclude that the established notion of computation is not adequate to describe Nature. Problems like this prompted the research community to investigate the construction of both *real* and conceptual computing devices capable of doing more than the established models of computation. This new field is known as *hypercomputation* (Syropoulos, Hypercomputation, 2008), however, we are not going to discuss these ideas in any detail.

THE SIMULATION HYPOTHESIS

The fascination about the abilities of mechanical devices has led many thinkers to ponder about the possibility that we actually live in a computer simulation. We have touched briefly this subject in the previous section but here we are going to elaborate on the idea and its pros and cons. Nick Bostrom (Bostrom, 2003) put forth his simulation hypothesis. More specifically, he argued that

- i. it is very probable that humanity will vanish before reaching a “posthuman” stage;
- ii. it will not be possible to run a significant number of simulations of the evolution of any posthuman civilization; and

- iii. it is very probable that we live in a computer simulation.

Trying to respond to such an argument is particularly difficult because it is similar to trying to convince a believer that there is no God. Even if we show that that no computing device, real or conceptual, can actually simulate physical systems, still one may argue that there are yet to be discovered computing systems capable of any kind of simulation.

SOLUTIONS AND RECOMMENDATIONS

Computers have been used to simulate physical systems for many years. However, there are some quantum systems that cannot be simulated by ordinary hardware and one should expect the wide-spread availability of quantum computers to be able to simulate such systems.

FUTURE RESEARCH DIRECTIONS

We presented work about non-computability of physical systems. The question whether computers (even *hypercomputers*) can actually simulate Nature is open. But another question is also open: Can computers simulate minds? The *Chinese Room Argument* by John Searle (Searle, 1980) has *shown* that computers can simulate minds but they cannot realize minds. This argument goes as follows: Imagine that Anna, who cannot speak or read Chinese, is locked in a room with boxes full of Chinese ideograms. In addition, she has at her disposal a rule book that enables her to answer questions put to her in Chinese. One may think of the rule book as a computer program. Anna receives ideograms that, unknown to her, are questions; she looks up in the rulebook what she is supposed to do; she picks up ideograms from the boxes, manipulates them according to the rules in the rule book, and hands out the required ideograms, which are interpreted as answers. We may suppose that Anna is able to fool an external observer, giving the impression that she actually speaks and understands Chinese. But clearly, Anna does not understand a word of Chinese. And if Anna does not understand although she appears to do so, then no computer will actually understand Chinese just because it is equipped with a computer program much like Anna's rule book. Thus, Turing computers cannot be intelligent.

CONCLUSION

We have discussed the notion of simulation of physical system by computers. We have also presented some techniques that use computer simulation to solve difficult computer programs. In addition, we discussed the limits of computer simulation and the simulation hypothesis. The conclusion is that simulation is a very important and useful tool but certainly not a panacea.

REFERENCES

Adami, C. (1998). *Introduction to Artificial Life*. Springer-Verlag. doi:10.1007/978-1-4612-1650-6

- Bostrom, N. (2003). Are You Living in a Computer Simulation? *The Philosophical Quarterly*, 53(211), 243–255. doi:10.1111/1467-9213.00309
- Dorigo, M., & Stützle, T. (2004). *Ant Colony Optimization*. The MIT Press. doi:10.7551/mitpress/1290.001.0001
- Feynman, R. P. (1982). Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6), 467–488. doi:10.1007/BF02650179
- Gardner, M. (1983). *Wheels, Life, and Other Mathematical Amusements*. W. H. Freeman and Company.
- Hayes, B. (2003). Debugging the Universe. *International Journal of Theoretical Physics*, 42(2), 277–295. doi:10.1023/A:1024447300346
- Joseph, I. (2016). *Step into a new world - Virtual Reality (VR)*. Retrieved 05 23, 2019, from <https://www.completegate.com/2016070154/blog/virtual-reality-explained>
- Kennedy, J., & Eberhart, R. C. (2001). *Swarm Intelligence*. Morgan Kaufmann Publishers.
- Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag. doi:10.1007/978-3-662-03315-9
- Osais, Y. E. (2017). *Computer Simulation: A Foundational Approach Using Python*. Chapman and Hall/CRC. doi:10.1201/9781315120294
- Pour-El, M. B., & Richards, J. I. (1989). *Computability in Analysis and Physics*. Springer-Verlag. doi:10.1007/978-3-662-21717-7
- Ringel, Z., & Kovrizhin, D. L. (2017). Quantized gravitational responses, the sign problem, and quantum complexity. *Science Advances*, 3(9), e1701758. doi:10.1126ciadv.1701758 PMID:28959729
- Romero, G. E., & Vila, G. S. (2014). *Introduction to Black Hole Astrophysics*. Springer. doi:10.1007/978-3-642-39596-3
- Searle, J. R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3(3), 417–424. doi:10.1017/S0140525X00005756
- Syropoulos, A. (2008). *Hypercomputation: Computing Beyond the Church-Turing Barrier*. Springer US. doi:10.1007/978-0-387-49970-3
- Syropoulos, A. (2017). *Demystifying Computation: A Hands-On Introduction*. World Scientific Publishing Europe. doi:10.1142/q0077
- Syropoulos, A. (2009). *Can we debug the Universe?* Retrieved 05 24, 2019, from <https://arxiv.org/abs/0910.2859>
- Thorne, K. S. (1994). *Black Holes & Time Warps: Einstein's Outrageous Legacy*. W. W. Norton & Company.

KEY TERMS AND DEFINITIONS

2

Computer Implementation: The creation of a computer system that exactly represents a physical a system.

Computer Simulation: A virtual representation of a physical system in silico that can be used to examine its properties.

A General-Purpose Uniform Random Number Package

E. Jack Chen

 <https://orcid.org/0000-0002-3038-8048>

BASF Corporation, USA

INTRODUCTION

As computer capacities and simulation (e.g., Monte Carlo method) technologies advance, simulation has become the method of choice for modeling and analysis. The fundamental advantage of simulation is that it can tolerate far less restrictive modeling assumptions, leading to an underlying model that is more reflective of reality and thus more valid, leading to better decisions (Lucas et al., 2015). Simulation studies are typically preceded by transforming in a more or less complicated way of a sequence of numbers between 0 and 1 produced by a pseudorandom generator into an observation of the measure of interest. Random numbers are a fundamental resource in science and technology. A facility for generating sequences of pseudorandom numbers is a fundamental part of computer simulation systems. Furthermore, random number generators also play an important role in cryptography and in the blockchain ecosystem. For example, common cryptosystems employ keys that must be generated in a random fashion. A collection of random variables $x_1, x_2, \dots, x_n$ is a random sample if they are independent and identically distributed. All samples of the sequence are generated independently of each other, and the value of the next sample in the sequence cannot be predicted, regardless of how many samples have already been produced.

True random numbers cannot be produced by a deterministic algorithm, and hence, random numbers generated by using a recursive equation are referred to as pseudorandom numbers. Pseudorandom numbers seem random, but in fact, they are not random and perfectly predictable. The purpose of these algorithms is to produce sequences of numbers whose behavior is difficult to distinguish from that of their “truly random” counterparts. The deterministic nature of these techniques is important because it can be reproduced effectively in computations. A facility for generating sequences of pseudorandom numbers is a fundamental part of computer simulation systems. Usually, in practice, such a facility produces a deterministic sequence of values, but externally these values should appear to be drawn independently from a uniform distribution between 0 and 1, i.e., they are independent and statistically indistinguishable from a truly random sequence. Furthermore, multiple independent streams of random numbers are often required in simulation studies, for instance, to facilitate synchronization for variance-reduction purposes, and for making independent replications.

A random number generator (RNG) is an algorithm that starting from an initial seed (or seeds), produces a stream of numbers that behaves as if it were a random sample when analyzed using statistical tests, see L’Ecuyer and Simard (2007) and Rukhin et al. (2010). The RNG is closely related to the Deterministic Random Bit Generators (DRBGs). See L’Ecuyer (1990, 2013) and references therein for more information on RNGs. We describe a portable set of software utilities for uniform random-number generation. It provides for multiple generators (streams) running simultaneously, and each generator (stream) has its sequence of numbers partitioned into many long disjoint contiguous substreams, see L’Ecuyer et al. (2002). Simple procedure calls allow the user to make any generator “jump” ahead/back

DOI: 10.4018/978-1-7998-3479-3.ch013

v steps (random numbers). This is more flexible than provides a pre-computed list of seeds that are pre-determined steps apart. Implementation issues are discussed. The basic underlying generator CMRG (Combined Multiple Recursive Generator) combines two multiple recursive random number generators with a period length of approximately 2^{191} ($\approx 3.1 \times 10^{57}$), good speed, and excellent theoretical properties, e.g., the lattice structure, see Kroese et al. (2011) for a list of desired properties.

BACKGROUND

There are several methods for generating the random numbers, of which the most popular are the congruential methods (mixed, multiplicative, and additive). These congruential methods are fast, convenient, and have good enough statistical properties when their parameters are chosen carefully. The (mixed) linear congruential generators (LCGs) are defined by

$$x_i = (ax_{i-1} + c) \text{MOD } m, u_i = \frac{x_i}{m}, x_0 \in \{1, \dots, m-1\}, i > 0.$$

where m (the modulus) is a positive integer (usually a very large primary number), a (the multiplier) $\in \{0, 1, \dots, m-1\}$ and c (the increment) is a nonnegative integer. This mathematical notation signifies that x_i is the remainder of $(ax_{i-1} + c)$ divided by m . Hence, $x_i \in \{0, 1, \dots, m-1\}$. Thus, random variable u_i is a uniform 0, 1 variable. Note that

$$x_{i+v} = \left(a^v x_i + \frac{c(a^v - 1)}{a - 1} \right) \text{MOD } m.$$

Hence, every x_i is completely determined by m , a , c , and x_0 . The sequence x_i repeats once it returns to a previously visited value. The period of a generator is the length of a generated steam before it begins to repeat. If $u_0 = u_p$ (where $p > 0$), then the length p is called the period. The longest possible period for a LCG is m , i.e., m represents the desired number of different values that could be generated for the random numbers. Hence, the modulus m is often taken as a large prime number close to the largest integer directly representable on the computer (i.e., equal or near $2^{31}-1$ for 32-bit computers). If $p=m$, we say that the generator has full period. The required conditions on how to choose m , a , and c so that the corresponding LCG will have full period are known, see Knuth (1997) or Law (2014).

LCGs are sensitive with respect to the parameters, especially the value of a . When $c > 0$, the LCGs are called mixed LCGs. When $c=0$,

$$x_i = ax_{i-1} \text{MOD } m, u_i = \frac{x_i}{m}, x_0 \in \{1, \dots, m-1\}, i > 0.$$

These LCGs are called multiplicative LCGs. Most existing implementations of LCGs are multiplicative LCGs, because in general the value of c does not have a large impact of the quality of an LCG. Note that if $x_i=0$, then all subsequent x_i are identically 0. Thus, the longest possible period for a multiplicative

LCG is $m-1$. Hence, to increase the period and to obtain a denser lattice which covers the space more evenly, one can increase the modulus m . Furthermore,

$$x_{i+v} = a^v x_i \text{ MOD } m.$$

Most experts now recognize that small LCGs with moduli around 2^{31} or so should no longer be used as general-purpose random-number generators. Not only can one exhaust the period in a few minutes on a PC (personal computer), but more importantly the poor structure of the points can dramatically bias simulation results for sample sizes much smaller than the period length.

One way of extending the basic LCG is to combine two or more LCGs through summation. Another way of extending the basic LCG is to use a higher-order recursion. A multiple recursive random number generator (MRG), which goes from integer to integer according to the recursion

$$x_i = (a_1 x_{i-1} + \dots + a_k x_{i-k}) \text{ MOD } m, u_i = \frac{x_i}{m}.$$

A seed $x_0, \dots, x_{k-2}, x_{k-1} \in \{1, \dots, m-1\}$, where i, k , and m are positive integers, and $a_1, \dots, a_k \in \{0, 1, \dots, m-1\}$. To increase the efficiency and ease the implementation, the MRG algorithm usually set all but two a_i 's to 0. Furthermore, the nonzero a_i should be small. However, L'Ecuyer (2013) points that these conditions are generally in conflict with those required for having a good lattice structure and statistical robustness. See Law (2014) on the lattice structure of pseudorandom numbers. The longest possible period for a MRG is m^k-1 (L'Ecuyer, 1996). Other way of extending the basic LCG is the additive congruential RNG (ACRON). The ACRON sets $a=1$ and replace c by some random number proceeding x_i in the sequence, for example, x_{i-1} (so that more than one seed is required to start calculating sequence). Wilkrmaratna (2008) indicated that the ACRON is a special case of a multiple recursive generator. Other classes of RNGs are available, e.g., twisted generalized feedback shift register generators (Matsumoto & Kurita, 1994; Matsumoto & Nishimura, 1998). See L'Ecuyer (2013) and the numerical algorithms group (2019) for more information on other methods of generating random numbers.

The RNGs discussed so far are by computational methods. Another approach is by physical methods, e.g., Kanter et al. (2010) and Raffaelli et al. (2018). Because the sequence of pseudorandom numbers is in principle predictable, data encrypted with pseudorandom numbers is potentially vulnerable to cryptanalysis. On the other hand, hardware random number generators produce sequences of numbers that are assumed to be unpredictable, and therefore provide the greatest security when used to encrypt data. Thus, this class of random bit generators is commonly known as non-deterministic random bit generators (NRBGs). Today, several physical devices are available on the market incorporate various types of methods to generate random bits or random digits. A comparative list can be found on Wikipedia (2019). Those hardware generators are needed for cryptography, games, lotteries, and in gambling machines, where true randomness is needed. Hardware random number generators generally produce only a limited number of random bits per second. In order to increase the available output data rate, they are often used to generate the "seed" for a faster pseudorandom number generator, which then generates a pseudorandom output sequence at a much higher data rate. These applications usually combine an algorithmic RNG with a small source to inject some true randomness, see, e.g., Crypto Poker Club (2018).

Hardware RNGs are not appropriate for simulation for many reasons. It was recognized already in the early 1950's that physical devices are complicated to install and run, costly, slow, and much more impor-

tantly cannot reproduce the same sequence repeatedly. Reproducibility is very important in simulation experiments, e.g., for comparing systems with common random numbers (Law, 2014) and for program verification and debugging. These drawbacks make it necessary for simulation to use algorithmic RNGs. L'Ecuyer (2010) points out that those physical devices are unnecessary and unpractical for Monte Carlo methods, because deterministic algorithmic methods are much more convenient than hardware devices.

MAIN FOCUS

We discuss the algorithm and implementation of L'Ecuyer's (1999) combined multiple recursive random number generator. A combination of generators can have a much longer period than any of its components. Furthermore, with well-chosen parameters, the CMRG has good structural properties and passes many statistical tests, see Kroese et al. (2011). The CMRG is one of the best general-purpose RNGs that are currently available. A list of classical LCGs and their properties can be found on Karl Entacher's website: <http://random.mat.sbg.ac.at/results/karl/server/>.

Basic Generators

The combined multiple recursive random number generator, which goes from integer to integer according to the recursion

$$x_{j,n} = (a_{j,1}x_{j,n-1} + \dots + a_{j,k}x_{j,n-k}) \text{MOD } m_j; \text{ for } j = 1, \dots, J,$$

$$Z_n = \sum_{j=1}^J (\delta_j x_{j,n}) \text{MOD } m_1$$

$$u_n = \begin{cases} z_n / (m_1 + 1) & \text{if } z_n > 0 \\ m_1 / (m_1 + 1) & \text{if } z_n = 0 \end{cases}$$

Where J is the number of MRGs used in the CMRG, j , n , k and m_j are positive integers, and each $a_{j,k}$ belongs to $Z_{m_j} = \{0, 1, \dots, m_j - 1\}$, δ_j are integers, and the greatest common divisor of δ_j and m_j is one for each j . To have fast algorithms, with a given j most $a_{j,k}$ should be 0. The transformation of u_n makes sure that u_n is never equal to 0 or 1 (otherwise, trouble may arise, e.g., when taking the logarithm to generate an exponential random variate). Assume that the m_j are distinct primes and that each recurrence j has a primitive characteristic polynomial, and thus period length $\rho_j = m_j^k - 1$ and the longest possible period for a CMRG is $\rho_1 \dots \rho_J / 2^{(J-1)}$ (L'Ecuyer, 2013).

The combined generator is a close approximation to a single MRG with a modulus equal to the product of the moduli of the components MRGs. Thus, the CMRG has the advantages associated with a larger modulus while permitting an implementation using smaller values. The CMRG combining multiple recursive sequences provides an efficient way of implementing random-number generators with long periods and good structural properties. If the parameters are well chosen, such generators are statistically far more robust than simple linear congruential generators that fit into a single computer word (L'Ecuyer, 1999).

The Need for Multiple Substreams

Many disjoint random number subsequences are often required in simulation studies, for instance, (1) to make independent replications and/or (2) to associate distinct “streams” of random numbers with different sources of randomness in the system to facilitate synchronization for variance reduction. To produce such “streams,” different seeds (values of vector $\mathbf{S}_i$) must be obtained far enough apart in the sequence to ensure that the streams do not overlap. Selection of seeds should also consider statistical properties between streams, such as apparent independence.

In other words, given any X_i (with seed $\mathbf{S}_i$) and positive integer v , there should be a quick way to compute X_{i+v} (with seed $\mathbf{S}_{i+v}$) without generating all intermediate values. The availability of efficient jump-ahead methods is very useful because it permits one to partition the RNG sequence into long disjoint stream and substreams of random numbers. Most packages offer no facility for jumping ahead directly from X_i to X_{i+v} or to compute distant seeds efficiently. Many simulation languages offer a limited number of streams, all based on the same generator, but using fixed starting seeds set say 100,000 values apart. This provides relatively low flexibility. Suppose, for instance, that you want to perform independent pairs of replications with common random numbers across the configurations (i.e., between any two runs of the same pair) to compare two different configurations of a system.

To insure proper synchronization, you want every generator to start from the same seed in both runs of the same pair. However, in general, these two runs will make a different number of calls to a generator, and programming “tricks” should be used to skip a proper amount of random numbers to resynchronize the generators for the next pair without overlap in the random number streams. This requires extra programming effort and could be error prone. Good software tools should ease the programmer’s task in that respect. A simple procedure call should permit resetting a generator to previous seed or jumping ahead to a new seed for the next run. Of course, the sequence of “new seeds” (one per run) should be the same of both configurations for the system. Implementing such tools requires efficient “jumping ahead” facilities, which in turn ask for efficient procedures to compute $(a \times s \text{ MOD } m)$, where a and s are positive integers. For examples showing the usefulness of substreams, see L’Ecuyer (2015).

Computing $(a \times s \text{ MOD } m)$

Consider a 32-bit computer on which all integers between -2^{32-1} and $2^{32-1} - 1$ are well represented. We want to compute $(a \times s \text{ MOD } m)$, where a , s , and m are positive integers smaller than 2^{32-1} . Without loss of generality, we assume that $a < m$ and $s < m$ (if not, replace a and s by $a \text{ MOD } m$ and $s \text{ MOD } m$, respectively). To keep seeds with 32-bit precision, all operations cannot produce any number greater than 2^{53} (the IEEE-754 standard, i.e., the floating-point numbers have at least 53 bits of precision for the mantissa). Therefore, special algorithm was needed to compute $(a \times s \text{ MOD } m)$, where a and s are less than 2^{32} (around 4.3×10^9).

The following algorithm was used. Let

$$a = a_1 \times 2^{17} + a_2$$

so

$$a \times s = a_1 \times s \times 2^{17} + a_2 \times s$$

Therefore,

$$a \times s \text{ MOD } m = \left((a_1 \times s \text{ MOD } m) \times 2^{17} + a_2 \times s \right) \text{ MOD } m$$

Where $a_1 < 2^{15}$, so $a_1 \times s < 2^{47} (= 2^{15} \times 2^{32})$, and $a_2 < 2^{17}$, so $a_2 \times s < 2^{49} (= 2^{17} \times 2^{32})$. Because $z = (a_1 \times s \text{ MOD } m) < 2^{32}$, we have $z \times 2^{17} < 2^{49}$ so that all the intermediate terms in the above computations are less than 2^{53} . Therefore, all seed values will have exact accuracy.

Note that $(a \times s \text{ MOD } m)$, where a and s are less than 2^{32} , can be computed directly on 64-bit computers that can perform operations with 128-bit quadruple precision, i.e., numbers as large as 2^{113} (around 1.0×10^{34}).

Computing the Jumping Matrices

The initial state of substreams can be computed easily if jumping-ahead facilities are available for the individual MRG components; that is, if an efficient algorithm is available for computing the state of the MRG v steps ahead of the current one, for large values of v . L'Ecuyer (1990) explained one way of doing that, based on the fact that the MRG can be viewed as a LCG in matrix form, whose state is a k -dimensional vector and whose multiplier is a $k \times k$ matrix A . To jump ahead by v values, just multiply the current state by $(A^v \text{ MOD } m)$. The matrix $(A^v \text{ MOD } m)$ can be pre-computed in time $O(\log_2 v)$, using the divide-and-conquer algorithm (Knuth, 1997).

That is, for the MRG random number generator, X_{i+v} can be computed directly from X_i using

$$X_{i+v} = (A^v X_i) \text{ MOD } m = (A^v \text{ MOD } m) X_i \text{ MOD } m$$

where

$$A = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ a_k & a_{k-1} & \cdots & a_1 \end{pmatrix}$$

is an invertible $k \times k$ matrix and

$$X_i = \begin{pmatrix} X_i \\ X_{i+1} \\ \vdots \\ X_{i+k-1} \end{pmatrix}$$

is a $k \times 1$ vector. When A has this special structure, the first $k-1$ components of X_i are obtained by shifting the last $k-1$ components of X_{i-1} , and the last component of X_i is a linear combination of the components of X_{i-1} , according to the MRG (L'Ecuyer, 1990). Kroese et al. (2011) point out that an MRG can be interpreted and implemented as a matrix multiplicative congruential generator, i.e.

$$X_i = (AX_{i-1}) \text{MOD } m, i = 1, 2, \dots$$

The divide-and-conquer algorithm (Knuth, 1997) was used to compute the jumping matrices A^v using the following recursion:

$$A^v \text{MOD } m = \begin{cases} A & \text{if } v = 1; \\ A(A^{v-1} \text{MOD } m) \text{MOD } m & \text{if } v > 1, v \text{ odd}; \\ (A^{v/2} \text{MOD } m)(A^{v/2} \text{MOD } m) \text{MOD } m & \text{if } v > 1, v \text{ even}. \end{cases}$$

A Random Number Package

We now propose a portable set of utilities for random number generation. We consider a basic underlying generator of period p . Let s_0 be the basic seed (initial state) for this generator and $s_1, s_2, \dots$ be its sequence of successive states. Let T denote the transition function of the generator, that is, the operator $T: S \rightarrow S$ such that $T(s_i) = s_{i+1}$, and T^q its q -fold composition ($T^q(s_i) = s_{i+q}$). Starting from states $I_1 = s_0$, $I_2 = s_q = T^q(I_1)$, $I_3 = s_{2q} = T^q(I_2)$, ..., $I_G = s_{(G-1)q} = T^q(I_{G-1})$, respectively. Each of these G sequences correspond to a “stream”, with length q . Note that q needs to be a very large number, say no smaller than 10^9 . Hence, there will be G virtual RNGs. I_g is called the *initial seed* of stream g . At any moment during program execution, stream g is in some state C_g , say, in its subsequence number r , that is, such that $C_g = T^{r-1}(I_g)$. We call C_g the *current state* of stream g .

The product of $a_{ij}(m_j-1)$ is less than 2^{53} . If this condition holds, then the integer $a_{ij}x_{ij}$ is always represented exactly in floating point on a 32-bit computer that supports the IEEE floating-point arithmetic standard, with at least 53 bits of precision for the mantissa. The generator can then be implemented directly in floating-point arithmetic, which is typically faster than an integer arithmetic implementation. On the other hand, with this implementation, the state of the generator is represented over $64k$ bits, as opposed to $32k$ bits when the x_{ij} are represented as 32-bit integers.

The mrg32k3a of L'Ecuyer (1999) implemented a CMRG with 2 components ordered 3, whose coefficients satisfy above condition. The moduli and coefficients are $m_1 = 2^{32} - 209 = 4294967087$, $a_{11} = 0$, $a_{12} = 1403580$, $a_{13} = -810728$, $m_2 = 2^{32} - 22853 = 4294944443$, $a_{21} = 527612$, $a_{22} = 0$, $a_{23} = -1370589$. Its period length is $\rho = (m_1^3 - 1)(m_2^3 - 1) / 2 \approx 2^{191} \approx 3.1 \times 10^{57}$. This implementation set $\delta_1 = -\delta_2 = 1$. The parameters have been chosen so that the period is long, a fast implementation is available, and the generator performs well with respect to the spectral test. The mrg32k3a has been implemented as one of the core random number generators in many simulation packages.

Before this procedure is called for the first time, one must initialize the global variables s_{10}, s_{11}, s_{12} to (exact) non-negative integer values less than m_1 and not all zero, and s_{20}, s_{21}, s_{22} to non-negative

integers less than m_2 and not all zero. The vectors (s10, s11, s12) and (s20, s21, s22) are the initial values of $(x_{1,0}, x_{1,1}, x_{1,2})$ and $(x_{2,0}, x_{2,1}, x_{2,2})$, respectively. They constitute the *seed*.

Therefore, in this implementation

$$x_{1,n} = (1403580.0x_{1,n-2} - 810728.0x_{1,n-3}) \text{MOD} 4294967087$$

$$x_{2,n} = (527612.0x_{2,n-1} - 1370589.0x_{2,n-3}) \text{MOD} 4294944443$$

$$Z_n = (x_{1,n} - x_{2,n}) \text{MOD} 4294967087$$

$$u_n = \begin{cases} z_n / 4294967088 & \text{if } z_n > 0 \\ 4294967087 / 4294967088 & \text{if } z_n = 0 \end{cases}$$

The matrix A_1 for s_{1n} to jump ahead one step is

$$A_1 = \begin{pmatrix} 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ -810728.0 & 1403580.0 & 0.0 \end{pmatrix}$$

and the matrix A_2 for s_{2n} to jump ahead one step is

$$A_2 = \begin{pmatrix} 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ -1370589.0 & 0.0 & 527612.0 \end{pmatrix}$$

The following is an implementation of the CMRG (L'Ecuyer's mrg32k3a) in C.

```
#define norm 2.328306549295728e-10 // 1.0/(m1+1)
#define norm2 2.328318825240738e-10 // 1.0/(m2+1)
#define m1 4294967087.0
#define m2 4294944443.0
// the initial seed
double s[2][3] = {
    {0.0, 0.0, 1.0},
    {0.0, 0.0, 1.0}
};
double MRG32k3a()
{
    long k;
    double p;
```

```

p = 1403580.0 * s[0][1] - 810728.0 * s[0][0];
k = long (p / m1); p -= k*m1; if (p < 0.0) p += m1;
s[0][0] = s[0][1]; s[0][1] = s[0][2]; s[0][2] = p;
p = 527612.0 * s[1][2] - 1370589.0 * s[1][0];
k = long (p / m2); p -= k*m2; if (p < 0.0) p += m2;
s[1][0] = s[1][1]; s[1][1] = s[1][2]; s[1][2] = p;
if (s[0][2] <= s[1][2]) return ((s[0][2] - s[1][2] + m1) * norm);
else return ((s[0][2] - s[1][2]) * norm);
}

```

See Chen (2016) for examples on how to use this RNG. The mrg32k3a RNG software packages are available for download via url <http://www.iro.umontreal.ca/~lecuyer/myftp/streams00/>

Jumping Backward

The jump-back matrix also exists. This is true because each component has a primitive characteristic polynomial (a necessary condition). Given a vector s_n , one may jump back v steps to s_{n-v} . For example, the jump-back-one-step matrix B_j is $A_j^{\rho_j-1} \text{ MOD } m_j$, where ρ_j is the period of the j th MRG of the CMRG. As one might expect, $A_j^v \times B_j^v \text{ MOD } m_j$ turns out to be the identity matrix I , for $v=1,2,\dots,\rho_j$. This equation is consistent with our intuition that $A_j^{\rho_j-1} \text{ MOD } m_j = I$, and this follows from Fermat's theorem. The matrix B_j generates the same stream but in reverse order. As we can see from the new recursion below, the values of the parameters b_{ij} are much larger than the original ones, where b_{ij} are the parameters of the new CMRG. Furthermore, we no longer have $b_{ij}(m_j-1) < 2^{53}$ for all b_{ij} . Therefore, the original recursion provides a more efficient generator.

The matrix B_1 for s_{1n} to jump back one step (i.e., jump ahead ρ_1-1 step) is

$$B_1 = \begin{pmatrix} 184888585.0 & 0.0 & 1945170933.0 \\ 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{pmatrix}$$

and the matrix B_2 for s_{2n} to jump back one step (i.e., jump ahead ρ_2-1 step) is

$$B_2 = \begin{pmatrix} 0.0 & 360363334.0 & 4225571728.0 \\ 1.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 \end{pmatrix}$$

The reverse stream follows the recursion

$$x_{1,n} = (184888585.0x_{1,n+1} + 1945170933.0x_{1,n+3}) \text{ MOD } 4294967087$$

$$x_{2,n} = (360363334.0x_{2,n+2} + 4225571728.0x_{2,n+3}) \text{MOD} 4294944443$$

or

$$x_{2,n} = (360363334.0x_{2,n+2} - 69372715x_{2,n+3}) \text{MOD} 4294944443$$

Generating Nonuniform Random Variate

To simulate the underlying system under study, various nonuniform random variates are required. These random variates are used to simulate the observations from the desired distribution. Random variates from distributions other than the uniform $[0,1]$ are generated by applying further transformations to the output value u_i of the uniform RNG. The simplest way of generating a random variate X with distribution function F is using the inverse transformation. Let $U(a,b)$ denote the uniform distribution between a and b . Based on the fact that $y=F(x) \sim U(0,1)$, we generate a random variate $U \sim U(0,1)$ and return the value $X=F^{-1}(U)$ as a variate with distribution F . Here F^{-1} is the inverse distribution function (or quantile function) and is defined by

$$X = F^{-1}(U) = \inf \{x : F(x) \geq U\}.$$

Note that

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

For more information of using inverse transform to generate random variates, see Law (2014). Other methods are available when F^{-1} is difficult or expensive to compute, see Kroese et al. (2011).

FUTURE TRENDS

Random number generators continue to be key components of any computer simulation, cryptography, and other areas where producing an unpredictable result is desirable. As computers become 64-bits based operation, RNGs based on 64-bits will be developed. Utilities for RNGs will be incorporated into many simulation packages and will be readily available for simulation practitioners. Furthermore, newer true random number generators based on hardware devices are continually being developed. These hardware devices are important components of blockchain environment, where cryptography and security are essential.

CONCLUSION

Sequence of pseudorandom numbers is not truly random, because it is completely determined by an initial value, i.e., seed. Although sequences that are truly random can be generated using hardware random

number generators, pseudorandom number generators (based on mathematical algorithm) are important in practice for their speed and their reproducibility. We have discussed the backbone CMRG random number generator, several utilities to enhance its practical use, and implementation issues. Our implementation eliminates the need to use a fixed number of pre-computed seeds, which would provide little flexibility. The proposed random-number-generation package provides jumping facilities, has good speed, a long period length, and excellent theoretical/statistical properties. It provides for multiple generators (streams) running simultaneously, and each generator (stream) has its sequence of numbers partitioned into many long disjoint contiguous substreams. The RNG mrg32k3a has been studied intensively and implemented in several software packages.

REFERENCES

- Chen, E. J. (2016). *Crafts of simulation programming*. World Scientific Publishing. doi:10.1142/9899
- Crypto Poker Club. (2018). *Crypto poker club's triple-factor random number generator (TFRNG) revealed*. Retrieved May 7, 2019, from <https://medium.com/@CryptoPokerClub/crypto-poker-clubs-triple-factor-random-number-generator-tfrng-revealed-4b83fccb7aea>
- Kanter, I., Aviad, Y., Reidler, I., Cohen, E., & Rosenbluh, M. (2010). An optical ultrafast random bit generator. *Nature Photonics*, 4(1), 58–61. doi:10.1038/nphoton.2009.235
- Knuth, D. E. (1997). *The art of computer programming*. 2 (3rd ed.). Addison-Wesley.
- Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). *Handbook of Monte Carlo Methods*. John Wiley & Sons. doi:10.1002/9781118014967
- L'Ecuyer, P. (1990). Random numbers for simulation. *Communications of the ACM*, 33(10), 85–97. doi:10.1145/84537.84555
- L'Ecuyer, P. (1996). Combined multiple recursive generators. *Operations Research*, 44(5), 816–822. doi:10.1287/opre.44.5.816
- L'Ecuyer, P. (1999). Good parameter sets for combined multiple recursive random number generators. *Operations Research*, 47(1), 159–164. doi:10.1287/opre.47.1.159
- L'Ecuyer, P. (2010). Pseudorandom number generators. In R. Cont (Ed.), *Encyclopedia of Quantitative Finance*. doi:10.1002/9780470061602.eqf13003
- L'Ecuyer, P. (2013). *Random number generation*. Retrieved May 6, 2019, from <http://www.iro.umontreal.ca/~lecuyer/myftp/papers/handstat.pdf>
- L'Ecuyer, P., Munger, D., Oreshkin, B., & Simard, R. (2017). Random Numbers for Parallel Computers: Requirements and Methods. *Mathematics and Computers in Simulation*, 135, 3–17. doi:10.1016/j.matcom.2016.05.005
- L'Ecuyer, P., & Simard, R. (2007). TestU01: A C library for empirical testing of random number generators. *ACM Transactions on Mathematical Software*, 33(4), 22. doi:10.1145/1268776.1268777

L'Ecuyer, P., Simard, R., Chen, E. J., & Kelton, W. D. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research*, 50(6), 1073–1074. doi:10.1287/opre.50.6.1073.358

Law, A. M. (2014). *Simulation modeling and analysis* (5th ed.). McGraw-Hill.

Matsumoto, M., & Kurita, Y. (1994). Twisted GFSR generators II. *ACM Transactions on Modeling and Computer Simulation*, 4(3), 254–266. doi:10.1145/189443.189445

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1), 3–30. doi:10.1145/272991.272995

Raffaelli, F., Sibson, P., Kennard, J. E., Mahler, D. H., Thompson, M. G., & Matthews, J. C. F. (2018). Generation of random numbers by measuring phase fluctuations from a laser diode with a silicon-on-insulator chip. *Optics Express*, 26(16), 19730–19741. doi:10.1364/OE.26.019730 PMID:30119294

Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., & Vo, S. (2010). A statistical test suite for random and pseudorandom number generators for cryptographic applications. *National Institute of Standard and Technology*. Retrieved May 7, 2019, from <https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-22r1a.pdf>

The Numerical Algorithms Group. G05 – Random number Generators. Retrieved May 6, 2019, from https://www.nag.co.uk/numeric/fl/nagdoc_fl23/pdf/G05/g05intro.pdf

Wikipedia. (2019). *Comparison of Hardware Random Number Generators* — Wikipedia, The Free Encyclopedia. Retrieved May 6, 2019, from https://en.wikipedia.org/wiki/Hardware_random_number_generator

Wilkrmaratna, R. S. (2008). The additive congruential random number generator – A special case of a multiple recursive generator. *Journal of Computational and Applied Mathematics*, 216(2), 371–387. doi:10.1016/j.cam.2007.05.018

ADDITIONAL READING

Barker, E. & Kelsey, J. (2012). *Recommendation for random number generation using deterministic random bit generators*. NIST SP800-90A.

Brent, R. P. (2007). Some long-period random number generators using shifts and xors. *The ANZIAM Journal*, 4, C188–C202. doi:10.21914/anziamj.v48i0.40

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods* (2nd ed.). Springer.

Goresky, M., & Klapper, A. (2003). Efficient multiply-with-carry random number generators with maximal period. *ACM Transactions on Modeling and Computer Simulation*, 13(4), 310–321. doi:10.1145/945511.945514

Haramoto, H., Matsumoto, M., Nishimura, T., Panneton, F., & L'Ecuyer, P. (2008). Efficient jump ahead for F2-linear random number generators. *INFORMS Journal on Computing*, 20(3), 385–390. doi:10.1287/ijoc.1070.0251

- Hörmann, W., & Leydold, J. (2003). Continuous random variate generation by fast numerical inversion. *ACM Transactions on Modeling and Computer Simulation*, 13(4), 347–362. doi:10.1145/945511.945517
- Hörmann, W., Leydold, J., & Derflinger, G. (2004). *Automatic nonuniform random variate generation*. Springer-Verlag. doi:10.1007/978-3-662-05946-3
- L'Ecuyer, P. (2017). History of Uniform Random Number Generation. *Proceedings of the 2017 Winter Simulation Conference*, 202–230. 10.1109/WSC.2017.8247790
- L'Ecuyer, P., Wambergue, P., & Bourceret, E. (2019). (to appear). Spectral Analysis of the MIXMAX Random Number Generators. *INFORMS Journal on Computing*.
- Lucas, T. W., Kelton, W. D., Sánchez, P. J., Sanchez, S. M., & Anderson, B. L. (2015). Changing the paradigm: Simulation, now a method of first resort. *Naval Research Logistics*, 62(4), 293–303. doi:10.1002/nav.21628
- Panneton, F., & L'Ecuyer, P. (2010). Resolution-stationary random number generators. *Mathematics and Computers in Simulation*, 80(6), 1096–1103. doi:10.1016/j.matcom.2007.09.014
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). Cambridge University Press.
- Wichmann, B. A., & Hill, I. D. (2006). Generating good pseudo-random numbers. *Computational Statistics & Data Analysis*, 51(3), 1614–1622. doi:10.1016/j.csda.2006.05.019

KEY TERMS AND DEFINITIONS

Blockchain: A blockchain is a growing list of digital records of transactions, called blocks, which are linked together using cryptograph. Each block contains a cryptographic hash of the previous block, a timestamp and transaction data.

Common Random Numbers: A variance reduction technique that uses the same random number streams to simulation different systems.

Deterministic Algorithm: An algorithm that given the same input always products the same outputs.

Deterministic Sequence: A sequence that the value of any future elements can be determined when the value of the current element is known

Jump Ahead (Back) v Steps: A utility to advance (go back) v steps in the random number stream without going through the stream sequentially.

Period Length (of RNGs): The length of the random number sequence before it recycles itself.

Quantile Function (Inverse Distribution Function): The quantile function (given a probability p) returns the value at or below which $100p$ percent of the population lies.

Random Number Generators (RNGs): Computational or physical devices designed to generate a sequence of numbers that appear to be random, as determined by statistical tests.

Variance Reduction: A technique to reduce the variance of the estimate, hence, increase the precision of the estimate.

Individual Cloud: After Cloud

2

Shigeki Sugiyama

 <https://orcid.org/0000-0003-3764-4160>

Independent Researcher, Japan

INTRODUCTION

It is just now at the top of an aggregation point of globalizations' era in terms of things and living creatures. And the communication methods including in many sorts of transfers like commodity, facility, information, system, thought, knowledge, people, culture, heritage, etc. may cause many kinds of and many types of interactions among us, which may have made the world closer but complex, smaller but intense, influential but solitude, networked but dividable, accumulated but distributed, direct but stand-still, stable but chaotic, abstracted but differentiated loosely, many possibilities but almost none, free but controlled, unique but ubiquitous, solitude but mass, directed but diverge, and accumulated knowledge but ordinal. As the results of these situations, they may cause even an individual world much wider or smaller, closer or divided, quicker or slower, dependent or independent, massive or nothing, unique or ubiquitous, far expanded or shrunk, any opportunity expected or nothing desired, viewable or invisible, manageable or uncontrollable, simple or chaotic, and limited or borderless. As it is observed, these phenomenon in the world is being existed as an extreme opposite twin (dipole, pair) unit at the same moment. This has been the causations of the Multi Dipoles. This phenomenon of the Multi Dipoles has caused the situations of many kinds of Mini Clouds by being used of the Information Technology (IT) for offering necessary information to us individually. This phenomenon of the Multi Dipoles seems to be further deepened and extended under this phenomenon. But on the contrary, this phenomenon turns out to be the opposite direction. That is to say, this phenomenon will make the Multi Dipoles converge into Unique Dipole in the very near future as the public usages will go to the two kinds of categories (groups) of people; Simple Users and Controlled Venders.

Talking about the functions of Mini Clouds in the same flow, it will become possible to access seamlessly to any Mini Cloud whose phenomenon will make the Mini Clouds converge into the Unique Cloud. Under these situations and conditions, an individual is able to get a thing and information as many and as much as being wanted in order to accomplish one's desire, which is very easy to do. But on the other hand, it is very true to say that the thing and the information will be offered to anybody as "a unit (a black box)" that it is not possible to go into further details to change, moderate, and manage a concerned matter. This phenomenon will likely lose a thing and information of originated matters silently without being recognized by users as many and as much as users used to have.

And it also can be said that, under these situations and conditions, almost everything in the world may interfere one another randomly, massively, closely, seamlessly, dividedly without being noticed to the users. And what is more, this phenomenon will go on further beyond year by year aggressively. In these chaos like conditions, everything and every matter will be gone further deepened Clouds entities inside, so that nobody will have a chance to recognize what is happening in reality world.

DOI: 10.4018/978-1-7998-3479-3.ch014

So here studies on the problems caused by these situations and conditions. And here will propose a method to resolve some of the problems caused by these situations by an idea of “Individual Cloud”.

BACKGROUND

There are so many kinds of tool, system, software, hardware, knowledge, thought, idea, principle, theory, and various kinds of human behaviors linked with Internetworking by Information Technologies that have been facilitating human actions even in cases at Dense Interference Space (DIS). DIS is a place where so many people or information is gathered randomly unintentionally on a networking or at a physical place. In the Interfering One Another (IOA) at DIS, it will be able to take any actions, for example, manufacturing fine goods, transporting goods, reducing a cost, solving a problem in managing a company, keeping an environment as they should be, making money, getting an information, making a new idea, learning, communicating, knowing, creating, diverging, etc. DIS could be appearing anywhere with any idea at any time purposely or coincidentally as a history tells us in the networked societies. And DIS may have an influential aspect in guiding a social and a technical direction without giving us any intention beforehand. And this tendency could not be stopped by anybody with any idea under the conventional ideas, thoughts, knowledge, and actions even though the tendency would be wanted to avoid.

In the present societies under these situations as mentioned above, there should be bearing various sorts of and many levels of DIS here and there. And some of them are brand new ones with a novel knowledge, and the others are mixed with new knowledge and conventional ones, and the another are with stuck of “obsoledge” (“obsoledge” means “obsolete” + “knowledge” by Toffler, 2010.).

Under this circumstance, any output of a service processing at DIS, for example, will be easily affected goodly or badly by Handling Method Of Knowledge (HMOK), Handling Method Of Knowledge Value (HMOKV), and Intelligence Of Knowledge (IOK) itself, which are closely related with cores of Clouds.

Technically speaking, it is now under “Techno National”, “Techno Hybrid”, and “Techno Global”. Under these situations, industries have been transferring their production spots to abroad from a developed country to reduce a production cost and a labor cost. As result of this, some of technologies of industries in those countries should be fading away and down to nothing (zero), and what is more, a deflation will be going on in those countries. Under these situations and conditions, ordinal machines equipped industries and traditional production methods used industries may have difficult and hard moments for surviving at the places where they used to have situated at. On the other hand, it is not so easy matter to create a new industry or a company in the replaces of the conventional companies as there is not any stable space for any purpose to be achieved easily in reality.

However, at present, in terms of these matter, it could be and can be thought of some of the methods to overcome these problems and the situations for surviving. And they are shown in the followings.

1. To become “Far Advanced Company in Technology and Management”:

This means that a company has the following functions.

- a. A company has a far advanced core technology.
- b. A company has the top share of distinguished goods in a limited field.
- c. A company has the best engineers and technicians in a field.
- d. A company has a unique management strategy and a good turnover.
- e. A company has constantly a new development.
- f. A company has collaboration with Universities.

- g. A company has good concerns with an environment and a natural resource.
 - h. A company may give employees nice motivations for working.
2. **To have “Alliance with Company in Developing Country”:**
This means that a company has a production spot in Asia or in South America in order to reduce a production cost and labor cost.
- a. A company has a distinguished management.
 - b. A company has a strategy of “Goods from Asia or from South America to the world”, for example.
 - c. A company has a unique planning section.
 - d. A company has a multiple management.
3. **To become “Global Company”:**
This means the following thing.
- a. A company will become a global company. That is to say, a company will stay at the best place (in any country) for production and management as multi-nationals.
4. **To become “Environment Concern Type of Company”:**
This means that a company has the core direction in sustainability in all.
- a. A company has the best concerns with the nature, the natural resources, and responsibilities for the goods produced.
5. **To become “Knowledge Accumulated Company”:**
This means the following things.
- a. Whatever we understand about the world, it is very hard to maintain an established position as a company for good. Generally speaking, it is understood that a company has to have an advanced thought comparing with others’. That is to say, it is very closely related with an idea how much knowledge, which is far advanced compared with others’, a company has. In this case, Knowledge is in a technical field and in a management field, which are closely related with Cloud.
6. **To become “Advanced Information Technology Equipped Company”:**
This means the following thing.
- a. A company has the advanced Information Technologies that may replace a human job by using Cloud.
7. **To become “Company with Niche Goods Production”:**
This means the following thing.
- a. A company has goods that might not interest in the major companies.
8. **To become “Company only in Net”: Company in Digital**
This means the following thing.
- a. A company does not have a physical office at all but it has all the necessary actions in Networking. This is a unique case of using Cloud as Knowledge.
9. **To become “Non Profit Organization (NPO)”:**
This means the following thing.
- a. An organization does not have a targeted making money motivation but has a loosely environment interfered action orientation with a nature, people, economy, culture, education, and the earth. This will be able to become a global but in the sense of economy mono pole.

Whichever the case to face to adopt, all these matters of 1 to 9 are closely related with “Cloud” for the necessary usages in planning, production, management, marketing, service, and sales.

PRESENT PHENOMENON IN SOCIETY AND PROBLEMS FACING

By taking an idea from 1 to 9 described in the previous chapter, a company may have any direction of the methods in order to keep its position in the world and in order to overcome the problems facing for being survived or for getting a higher position with shares and sales. However generally speaking, it is very true to say that the economic situation in the world is not vivid like many decades ago. What is worse, the consumers do not have much intention to buy goods and foods, etc. comparing with the decades ago. Because almost everything can be got in the world when it is desired through Networking, for example. They can be easily transported even if they are away from us in distance. They also can be simply transferred through a computer networking if they are data, images, and information technology related matters. They are able to come upon to us by the transfers without any information contained just as a matter of “THING”, which means that Goods has been arrived upon us just as a final matter of “THING” without making us known its Contents, Relations, and Meanings.

If it will be allowed to describe the above phenomenon with a bit of over expression, they may have lost these Contents, Relations, and Meanings of Goods and Foods before they have been delivered to us. And it is quite natural and easy to lose the values of them too at the time when they are used or are eaten. As the result of this, it is quite natural that the interests on Goods or Foods might be lost too although the producers and raisers have been doing much effort on adding the values onto them.

In another word, if it will be concluded to say, the world will be surrounded by full of simple matters of “THING”, which might make us easily loose interests on Goods and Foods. And ultimately, the world will be surrounded by “simply a function” and “simply a matter” by “THING” that have been asked to possess. As the result of this, a consumer will be fed up with them sooner or later by “a function” and “a simple matter”. And a new Goods or Foods will be asked for soon after by a consumer. And in this vicious circle, it might be said that a deflation will be on, which will make an economic basic ground become smaller and smaller and will be likely shrunk to zero without any notice to us.

So, in short, what are the problems under this phenomenon in the case of production, market, and sales? They will be the followings.

1. A consumer will easily loose an interest in Goods, etc. sooner or later, and so the values that are existed on Goods, etc. will be fading away to nothing suddenly and instantly. As the result of this, a market size will have a tendency to diminish itself suddenly and instantly.
2. DIS will come up to the world suddenly and without any notice. So a new market will have a tendency to come up to the world suddenly without any notice to us.
3. Tendency is not a matter that is able to control easily.
4. A value is not a matter that is able to create easily.
5. Consumers are independent nowadays, and so they are independently taking any thought and action in all over the world. As the result of this, to grasp and to know a mass behavior is not an easy matter.
6. Data processing methods are in full for use today but it is hard to use them properly. It is asked to have a method to use them simply, intentionally, knowledgably, and intelligently. This situation is really the matter of “Cloud” to come up and to use.

KEY FACTOR IN THE PROBLEMS SOLVING

In the previous chapter, the problems facing are extracted and are discussed. And here they will be studied further in order to find the core matter to cause the problems if there is any.

It will not take much time to find the core matter of the problems studied in the above chapter when it is thought out.

A new product has been asked almost every instance in commodities, in hobbies, and in lives. Or in another expression, a company has been offering a new product at every instance. This phenomenon shows that values, meanings, and relations of them have been lost sooner or later, which were existed on them. But what happens if those things likely loose values, meanings, and relations of them in much longer time. In another word, if much important and vivid information are kept on a certain goods, they will be kept for them until the mechanical functions are damaged for use. In another word, “Contents (C)”, “Relations (R)”, and “Meanings (M)” of Goods, for example, are the core matters (the core factors) for us to keep our interests on Goods. It also can be concluded by looking back of this phenomenon that those core matters (C, R, and M) should be strongly combined with “an intention to buy”.

It is obviously to say that the following two subjects of 1 and 2 are important actions and functions if there is a sale of Goods on a market;

1. Goods should have good reasons of “Contents”, “Relations”, and “Meanings” that belong to its own.
2. “Contents”, “Relations”, and “Meanings” should be within a quite understandable situation for anybody. And so, those things should be seen or taken or used quite easily by “anybody” at “any time” when those are wanted to.

Today, there are the following systems and functions in order to produce Goods in a costless and in a labor redundant manner as a broad sense of a management field.

SAP; (Enterprise Resource Planning)

SOA; Service Oriented Architecture

SaaS; Software as a Service

Mash Up; Mashing Up software

Meta Frame; (peer to peer relation)

PaaS; Platform as a Service

HaaS; Hardware as a Service

IaaS; Infrastructure as a Service

Cloud; Cloud Computing within a particular area

Mini Clouds; Various Kinds of Small Cloud

Those systems and functions mentioned above have been concluded as a part of “Cloud” computing in order to solve any kind of problem.

But nowadays there is the fact that it is not possible to invest so much money on a company system for solving the problems. Because skilled and educated engineers might be offered in the areas of developing countries, and those won’t cost so much comparing to the systems mentioned above. And they may be creative which cannot be done by the computer systems mentioned above. So the companies try to use those engineers instead of using a high cost of the computer systems with SAP, SOA, SaaS, PaaS,

HaaS, IaaS, Cloud, etc. And what is more, it is quite hard to create and add the two subjects mentioned above on Goods, for example, by using the above computer systems and the functions instantly.

CLOUD ORIENTED SERVICE IN SERVICE SCIENCE

Here introduces a new idea of “CLOUD ORIENTED SERVICE IN SERVICE SCIENCE” by using the method of “Accumulation and Integration in Seamless Knowledge”, which will be able to solve those problems mentioned in a sense of “Service” with a low labor cost and a low production cost.

DEFINITION OF CLOUD ORIENTED SERVICE BY BEHAVIOR

“By entering into the “CLOUD ORIENTED SERVICE” system, the desire (subject, problem) will be solved and offered by means of SERVICE; BPOSS, POSS, BFOSS, BSOSS, and IMOS. And this SERVICE consists of Field Analysis, Categorical Analysis, Functional Analysis, and Knowledge Enhancement in the mixed ways. And the services will be mainly offered by the computer systems, which may need some simple data-input.”

In this manner, Service as Field Analysis will be firstly defined with five service systems as described below.

Service

Service Field

1. Business Plan Oriented Service (BPOS):

Generally speaking, there are so many businesses in this world and also there will come up new businesses to the world day by day. It is true to say that each business itself is slightly different from one another in a sense of goods, a shape, a production method, a market, a customer, an environment, etc. But they may be divided as field (gathered as a bundle of groups) by a system structure, a kind of goods, a process kind, a market kind, etc., which should make the number of the businesses' fields small in quantity. That is to say, it will be possible to have a business plan for each field as “A Generalized System of Service” (Business Plan Oriented Service System: BPOS), which will be able to offer the following Service.

- a. BPOS can make use of database for a new generalized business plan (Contents).
- b. BPOS is able to offer one of the business plans for planning and controlling, or for getting assistance out of BPOS in managing.
- c. BPOS may make use of the database for using other purposes (Relations) by connecting to other systems with IT facilities like Printer, CAD/CAM, etc.

2. Project Oriented Service (POS):

Now it is a common thing that we have some projects derived from a public sector (government, or NPO, etc.), a private sector, and an Academy Sector. These are Project Oriented Service (a theme of a project purpose is defined), which means that these are usually processed by collaboration among Private

Sector, Academy Sector, and Public Sector. That is to say, each sector should be responsible for some parts in the project in order to get a purpose accomplished. In this case, the things that are needed will be decided before starting a project. So the services offered for projects will be the followings.

- a. POS can manage the whole project and can do “Check – Plan – Do – Evaluation - Action” circle. And this can offer and create Meanings.
 - b. POS has flexibility for adding and reducing a process of the system targeted.
 - c. POS may have a human interaction (of a coordinator) from time to time. In this case, a purpose of a project will be clearly decided before the start but not for methods to get the output.
3. Business Fostering Oriented Service (BFOS):

When a business plan starts, it is a usual thing to say that there might be a problem for processing at some part during the operation. This kind of problem should be sort out by as soon as possible. For doing this kind of problem solving, it is very useful to have a similar kind of process example, which will give and offer a solution tips for the problem facing. So the services offered will be the followings. After having gone through this process, a new knowledge will be created, or value added knowledge will be made on.

- a. BFOS can offer necessary information (Contents, Relations, and Meanings) including a method from the database accumulated from day by day processing when the business has a problem.
 - b. BFOS can offer a new-targeted mass (Relations) for sales, a new strategy for processing, and a new category for promotion, etc. by using an information retrieval system and GPS system.
 - c. BFOS can offer a market research activity (Relations) by using the several analysis methods offered by the BFOS.
4. Business Skill Oriented Service (BSOS):

There are many skills and tactics in a field of business when a business is under operation. And these skills will be usually stored in a computer system as a knowledge base in order to use them for future. And an old knowledge will be transformed into a new knowledge through this process.

- a. BSOS can offer business skills and tactics (Relations).
 - b. BSOS can store the skills and the tactics by the system in Database.
5. Information Management Oriented Service (IMOS): (Suzuki 2013)

It always happens to get new information, and to process it, and to accumulate it by adding other related information. And those of many kinds of accumulated information need to be managed in the way to facilitate a problem solving and desired structure formation in processing.

- a. IMOS can offer a problem solving technique by using various kinds of information on one computer screen. And those kinds of information can be searched, accumulated, processed, and added onto a system as a new knowledge by using Web., Google, and Data stored.
- b. IMOS can offer the publication processes in any stage. And also this can offer an increment in knowledge enhancement.

Here explains about the system behavior of “CLOUD ORIENTED SERVICE” system.

The system consists of the four categories (BPOS, POS, BFOS, BSOS, IMOS) as mentioned above. The processes for Service Action as Categorical Analysis will be explained in the following subsection.

Service Action:Sugiyama, Suzuki (2012)

1. Enter with “Subject”

Go to the system with “a subject”. And go further into **BPOS**, which will offer a related category with the subject. And find the categories related with the subject.

2. Get Subject Required

Find the category related with the subject that will offer a desired output.

3. Processing with Service

After having found the related category, it will start processing for inputting the subject in order to get a desired output.

4. Evaluate Output

It is very important to know whether the output offered is valuable/acceptable or not. If the output is all right for use, this will be terminated.

5. Goto First Stage

If the output is not valuable for use, go to the “first stage of the system” to find more appropriate category.

The processes for Service Improvement as Functional Analysis will be explained in the following subsection.

Service Improvement

1. Get Unsolved Subject

Get an unsolved subject.

2. Resolve It with Services

Find the related functions for the subject by using the related categories. And then by using the categories, find out relatively influenced functions by the subject. If there are adding factors and functions, manipulate them by using this system. And then store it as a generalized system for service.

3. Evaluate It with Output

Find out the output whether it is good for use or not. It is good for use, and then stores it as one of the services onto Service Category with the functions used.

4. Store It as Service

If the output is not valuable for use, go to the “first stage of the system”.

The General System Behavior

By using the method studied in the above sub chapters, we can create “**CLOUD ORIENTED SERVICE**” system of which the general system behavior is described below.

The system INPUT is for the desire at the place for “Service with Subject in a certain field”, and then the desire will be categorized according to the content (elements) of the desire. And then the process center will be transferred to “Mini Cloud” of which is the most suitable for the desire. And then “Mini Cloud” will be getting an appropriate answer by interchanging the information with “**CLOUD ORIENTED SERVICE**”. And then, the general behavior of this system will give us a notice what sort of service field and category that we are interested. By then, necessary functions will be ready for processing. By telling the category which we need to system, the system lures us one of Mini Cloud, which will give us an answer as OUTPUT.

However, there should be one more thing to think about regarding this matter. That is to say, “Whether or not that there is any limitation in Cloud?” Or “Mini Cloud is good enough for use ?”.

In the next chapter, the limitation will be considered by using a mini Cloud.

SIMPLE EXAMPLE; GSHOP AND INFORMATION MANAGEMENT ORIENTED SYSTEM

Gshop Structure in Field, Category, and Function

Here shows a simple example for making the Mini Cloud system by using Gshop in order to show limitation if there is any.

The whole system consists of the major five subsystems.

1. “Process Template” is to show the present processing contents and the processing results.
2. “ScriptEngine” is for various kinds of Data processing, like “print, mail, SNS, etc.”, “Search and Recommend by using GIS”, “various analysis”, etc.
3. “Database” is for Information, Object, and Person.
- 4 “Various Field Definitions”
 5. “Project Plans and related”

The system behavior for the processing will be shown below.

1. First Stage of the processing:

From “Analysis Threshold”, the system understands which field this subject is related with by using the “Evaluations”. And the system will find that this subject is related with “Project B” mostly. And at the same time, the system knows that “Project B” consists of “Processes A to E”. Each “Process” will get Data from “ScriptEngine” one by one. And the system will give necessary outputs.

2. Second Stage of the processing:

If the outputs are not as suitable as expected, then change the parameters on Process A and Process D. And get the outputs. If the results were good enough to proceed in the real world, then this system ceases to close.

3. Third Stage of the processing:

If the outputs are not as suitable as expected, then change the Processes Methods to Process E and etc. provided. And get the outputs. If the results were good enough to proceed in the real world, then this system ceases to close.

4. Fourth Stage of the processing:

If the outputs are not as suitable as expected, then try to search a new processing method by using a networking, from books, and with computer systems. And this processing will be able to add a new knowledge onto this system as shown in Figure (2) Seamless Knowledge in Gshop. The Knowledge in this processing is to add a new technical processing, a new data processing, a new manipulation, a new thought, a new behavior, a new transformation, and a new relation.

That is to say, this is a processing route to find a new knowledge/rule by abstracting the methods used here repeatedly.

5. Fifth Stage of the processing:

The information and the knowledge that will be accumulated in the Fourth Stage will be able to be enhanced, comprehended, inferred, and enhanced in order to fulfill the information and the knowledge.

These processes from 1 to 5 will be repeated endlessly until one is able to get the output satisfied.

By this method in the seamless knowledge creation processes, one can create a new project which will give a new plan with a new knowledge.

However, it is easy to see that it is very hard to moderate the output after the processes have been completed.

Information Management Oriented System (IMOS): Suzuki (2013)

Here shows a simple example for making Mini Cloud system by using Information Management Oriented System that there is any limitation for use.

This system consists of the following structures.

1. Subject (category);

Each subject consists of project. And each project consists of Contents, Categories, and Objects. Each one of them (Contents, Categories, and Objects) has a seamless relation in the project. So a project is able to be directed onto a certain purpose in order to accomplish a desire by accumulating, enhancing, managing, restructuring, and concluding.

2. Process (Link and Database)

Each process consists of Link and Database. Link and Database have related data, searching data by related data, enhanced knowledge, related data (document, Web, geographical data, XML data, etc.) And each process has a seamless relation in the process.

Through these processes, it is possible to make and enhance “knowledge” in **Mini Cloud**.

However, Mini Cloud may behave only in a limited space which has been created by the above process. This intelligence of Mini Cloud goes under “static behavior” and “closed space (not open for public)”.

AFTER CLOUD

From the above chapters, it is quite natural to see the following things.

1. Things and living creatures are transformed into “a digital world”.
2. The digital world consists of a world of “data”.
3. It might be said that the digital world is “a virtual world”.
4. The virtual world is “a simulation world”, which is not real.
5. “A hypothetical world” mostly leads to the simulation world.
6. It might be said that Cloud goes under the hypothetical world.

If Clouds are “static” and “closed” for users, the output given by Clouds will be the rigid ones. And so, they are all from the virtual world that Cloud has made them by itself. Under this circumstance, the space that the world is facing is absolutely a virtual world. The space must not be cared about whether or not the world is real or not.

There are many kinds of and many types of Mini Clouds in the world. They are individually existed for a targeted customer. But Mini Cloud is Mini Cloud that is able to foster in a very limited category or within a closed area. And it can be said that users have to use many kinds of and many types of Mini Clouds in order to accomplish ones desires. In another word, it has come to the time which all Mini Cloud in the world may act as “knowledge of the world” uniquely. At this time, it will be necessary to access each related Mini Cloud in order to construct “each own knowledge for use”. This is able to be done by the idea of Dynamic Knowledge Wisdom (Sugiyama 2013).

This behaves like “knowledge enhancement cloud” by Net Surfing with the data taken by the structure of groups of URL and groups of Mini Clouds with processing and restructuring.

Dynamic Knowledge Wisdom is a kind of “Smart Cloud (SC)”, which will get and accumulate necessary information by Net Surfing on Webs (Mini Clouds) one after another. SC is formed as series of Mini Clouds (MC), those related data, and those related knowledge as described below.

SC = [MC, Data, Knowledge, Hypothetical World| MC1, MC2, ..., MCn, Data, Knowledge 1, Knowledge 2, ..., Knowledge n, Simulation World 1, ..., Simulation World n]

Inside of Cloud content is searched basically by Behavior of State Comprehension and State Inference, which might be an extension of Eigen Vector and Eigen Value behaviors. And this Mini Cloud (MC) under these conditions is called “Individual Cloud”.

CONCLUSION

Here argued about the following things.

The present real world is equal to the digital world offered by group of Mini Clouds. The digital world is equal to the virtual world offered by Clouds. The virtual world is equal to the hypothetical world that will be made by knowledge computer and the humans. And the hypothetical world is equal to a kind of Cloud.

So those can be expressed as follows.

“The Present Real World” = “The Digital World” = “The Virtual World” = “The Hypothetical World by Simulation” = “Clouds; Mini Cloud, Unique Cloud, Smart Cloud”.

And the hypothetical world is able to be manipulated by a brand new method of “Accumulation and Integration in Seamless Knowledge” by using the knowledge processing mechanisms of “CLOUD ORIENTED SERVICE”, so called Smart Cloud, within a principle of Service Science.

And the following results have been given.

- a. It is shown that “SUBJECT in demand and need” will be resolved by means of the idea of “CLOUD ORIENTED SERVICE; BPOS, POS, BFOS, BSOS, and IMOS in a sense of “Service Science” with the knowledge processing mechanisms of “Accumulation and Integration in Seamless Knowledge”.
- b. By using this method, it will be quite possible to transform “Obsoledge” into “Brand New Knowledge”.
- c. HMOK, HMOKV, and IOK are able to be done by using this method.
- d. It is obvious that this method can contribute to “Cost Reduction and Labor Savings”.
- e. This method can offer any type of fostering to an individual by accumulating and integrating knowledge seamlessly that is likely creating a brand new knowledge.
- f. By using this method, an individual is able to do anything by oneself from the first stage to the final stage for a desire accomplishment assisted by Gshop.
- g. As a next phenomenon in Clouds, an idea of Dynamic Knowledge Wisdom as Smart Cloud is introduced as “Individual Cloud”.
- h. Individual Cloud will have an ability and knowledge to escape from the hypothetical world.

FUTURE WORK AND DIRECTIONS

It will be thought for sure that any kinds of knowledge will be accumulated and enhanced as Unique Cloud and Smart Cloud. And as a result of this phenomenon, the related space for the world will be able to be indulged in the world of getting processed information as a final use of data for a daily life. This is a very comfortable world to live, but, on the other hand, there might be a possibility which there will be many or some who live only under the offered information and knowledge from intelligent computer systems (Mini Cloud, Unique Cloud, Smart Cloud).

So, it is very important to have knowledge of us in investigating whether or not Unique Cloud and Individual Cloud are useful enough to use by inspecting from the outside world even though we are to live in the hypothetical world. Or, in another words, whether or not an individual wants to live in its own individual hypothetical world, or an individual wants to live in another's hypothetical world?

REFERENCES

- Sugiyama, S. (2008). Fundamental Behavior. In *Communication Method*. Beijing, China: IEEE/SOLI.
- Sugiyama, S. (2013). *Intangible Capital Management Method As Dynamic Knowledge Wisdom*. In *Intellectual Capital Management For Knowledge-Based Organizations*. IGI Global.
- Sugiyama, S., & Suzuki, J. (2012). *Accumulation and Integration in Seamless Knowledge*. *International Journal of Asian Business and Information Management*.
- Suzuki, J. (2009). *Gshop, Users' Manual*. PITS.
- Suzuki, J. (2013). *Multi Publications, Users' Manual*. PITS.
- Toffler & Toffler. (2010). *40 for the next 40*. Toffler Associates. Retrieved October, 2010, from www.htmq.com/mirai/next40.shtml

A Collaborative Ranking Approach for Discovery and Selection of Cloud Services

Maya Rathore

 <https://orcid.org/0000-0002-3452-4684>

Department of Computer Science, Christian Eminent College, Indore, India

Ugrasen Suman

 <https://orcid.org/0000-0001-9236-9232>

School of Computer Science and Information Technology, Devi Ahilya University, India

INTRODUCTION

Cloud computing is gaining dominance over the last few years along with its various features such as, on-demand ease of access of computing resources as well as software services. Computing facilities are designed and utilized as a service using virtualization techniques and used as automated business logic in both public and private sectors. Along with various benefits data privacy, data confidentiality and trust establishment are considered to be the main security concerns for an organization to move its data to the cloud platform (Xu et al., 2015; Elfirdoussi, Jarir & Quafafou, 2012). The hasty developments of cloud computing means that cloud services have become the main computing mode on the Internet. Numerous services are offered over cloud to offer analogous functionalities. Despite the fact, the difficulty in identifying trustworthy services has fascinated the attention of researchers (Xu et al., 2015; Elfirdoussi, Jarir & Quafafou, 2012). For this reason, the concepts of trust and reputation had been brought to assess the reliability of cloud services (Li, Zhou, & Yang, 2012).

Reputation is a prejudiced assessment of a cloud service based on personal experience of individual or the advocacy of other users. In recent times, a variety of reputation systems have been proposed to deal with the challenges posed by open and dynamic cloud service environments (Itani, Ghali, Kayssi, & Chehab, 2014). The focus of majority of these systems is on computation of reputation ratings, reputation management, experience, and other features of dynamic environments and provides an appropriate solution to users (Malik & Bouguettaya, 2009; Rathore & Suman, 2013a, 2013b, 2013c, 2016; Yuan, An, & Wang, 2009; Trang, Zhao, & Yang, 2010). However, the survival of biased ratings affects the accuracy of trust evaluations to a great extent. At present, the focus of these reputation models are mainly on the accuracy of trust evaluations (Wang, Zheng, Sun, Zou, & Yang, 2011; Malik & Bouguettaya, 2009a, 2009b); however, these existing methods are limited by personality preference.

At present, to deal with the issue of trustworthiness, computation of the reputation is one of popular method (Elfirdoussi, Jarir & Quafafou, 2012). Reputation is the subject of a lot of interdisciplinary research and can be defined as “the collected and processed information about an individuals’ behavior as observed by others” (Wang, Chao, Lo, Lin, & Wang, 2011). The user can trust the service to a certain extent, based on aggregated feedback given by earlier consumers can be considered as the reputation score of a cloud service. A common attribute in existing cloud computing models is that a universal reputation score is designed through these reputation systems and the value is the same for all consumers. It is commonly used in profit-making location, such as eBay’s feedback mechanism or Epinions (Li, &

DOI: 10.4018/978-1-7998-3479-3.ch015

Wang, 2011). However, in general the reputation is potentially subjective. Feedback ratings are based on the individual's usage experience of a cloud service, which always needs to be seen in the context of the users experience and expectations of a service.

Consider for example two cloud service consumers A (denoted as CA) and B (denoted as CB). They are both looking for a cloud storage service. An industry expert has extremely precise requirements for confidentiality. CB is an amateur whose main aim is to obtain cheap storage. Now consider CB a free storage service is expected to be extremely glad as the cost is minimal and if usability is ok he/she will rank this service very highly. CA might read the small print regarding data policies and is likely to find issues that are not to her liking; so they would rank the service somewhat lower. Of course both are correct in their individual rankings. A global reputation score would not reflect the service users accurately.

Recently, extensive research work has been carried out on reputation prediction of cloud service in which Collaborative Filtering (CF) is extensively used (Xu et al., 2015; Wang, Chao, Lo, Lin, & Wang, 2011; Itani, Ghali, Kayssi, & Chehab, 2014). The classic procedure of CF is firstly to recognize a identical users set or services with identical ratings based on Pearson Correlation Coefficient (PCC), and secondly to use the similar users ratings and/or services to make predictions. Historical feedback ratings contributed by different cloud users are the basis for this prediction. Therefore, the prediction accuracy of CF approaches is highly influenced by the trustworthiness of the user-contributed feedback ratings. However, the present CF approaches uses a theory that all consumers-contributed feedback ratings are trustworthy (Xu et al., 2015; Wang, Chao, Lo, Lin, & Wang, 2011). However, a few feedback ratings can be unreliable due to certain reasons in reality: 1) some users may always give the maximal/ minimal values for the services, (2) cloud service consumers who may be at the same time cloud service providers, intentionally give higher feedback ratings to their own services and terrible mouthing to others services. Therefore, it is significant to consider data credibility to enable more robust service QoS value prediction.

Most existing reputation computation approaches compute service reputation based on QoS parameter and service level. These approaches have a high degree of dependency on the service user's feedback ratings services (Li, Zhou, & Yang, 2012; Li, & Wang, 2011; Trang, Zhao, & Yang, 2010; Malik & Bouguettaya, 2009a, 2009b; Wang, Zheng, Sun, Zou, & Yang, 2011). Assessment of service reputation at QoS parameter level requires that each QoS parameter should be monitored and tracked dynamically because QoS may change over time. Reputation assessment at service level based on a user's feedback rating also bears many problems, such as bias towards positive feedback ratings (Malik & Bouguettaya, 2009a, 2009b). Another problem with user's feedback rating is to ensure its accuracy and trustworthiness that users provide without even knowing much about the service (Wang, Zheng, Sun, Zou, & Yang, 2011; Le-Hung, & Aberer, 2009).

In order to address aforesaid weaknesses, a collaborative ranking approach along with a cloud service reputation model is proposed in this book chapter for detecting the malicious users' feedback rating and evaluating the service reputation. The proposed approach computes the service reputation to discover and select best services over cloud with the combination of user's rating and access rate (Yuan, An, & Wang, 2009; Rathore & Suman, 2013a, 2013b, 2013c, 2016). The former is dependent on user's feedback which is reported after the invocation of services. The later is the access rate, which is obtained through active monitoring of every service to be invoked by service consumer. Malevolent feedback rating using similarity computations are detected by presented approach and reduces the effect of user feedback preferences by Pearson correlation coefficient. An algorithm is also presented for the proposed approach. Wide-ranging experimental result shows that the presented approach can be helpful in improving the reliability of cloud services to be discovered.

This book chapter is organized as follows. Necessary background details related with cloud service reputation is discussed in Section 2. Section 3 highlights the focus of the book chapter along with proposed solution. The proposed collaborative ranking approach along with cloud service reputation evaluation model is presented in Section 4. Section 5 shows the experimental analysis and results of proposed approach. Section 6 discusses the proposed work with its pros and cons and also provides the future research directions. Finally, Section 7 concludes the paper along with the future work.

BACKGROUND

The primary issue among the online as well as offline business communities has always been a reputation. The cloud service consumers choose cloud services based on their reputation. Reputation is a word that has diverse meanings in various domains. It is defined as a needle of whether a user is ready to select a service based on the evaluations of other users. Thus, the result of a reputation evaluation will affect the decision about whether to interact with a service provider. Feedback related to previous interactions among cloud service users and service providers are composed by a reputation system to predict its future reliability (Li, Zhou, & Yang, 2012). Several research works have been carried out on trust and reputation using various mechanisms, such as incentive based, association based, previous invocation based etc.

Selection and discovery of cloud services for individual task is performed with the help of a variety of Quality of Service (QoS) criteria such as reputation, access rate, etc. Cloud service reputation represents a general opinion about its goodness that has been rated based on its various features. It is a communal user's feedback rating who have interacted with or used the cloud service earlier. Feedback rating is the observation of each user about services that has been invoked by them. It could be a sole value in lieu of an overall opinion or a vector representing a value for each QoS parameter of a service, such as response time, reliability and availability. A framework for trust management is introduced to distinguish reputation for various user groups (Xu et al., 2015). One possible solution to achieve is to compute the similarity of consumers rating based on a decision-tree model which clusters feedback into localized scores. A time decay feature appropriate for feedback is also to be considered to refine the result.

Wide-ranging of research work has been carried out on prediction of cloud service reputation, such as collaborative filtering (CF), referral chain, contribution based and previous QoS historical records (Li, Zhou, & Yang, 2012; Li, & Wang, 2011; Itani, Ghali, Kayssi, & Chehab, 2014; Wang, Chao, Lo, Lin, & Wang, 2011). Trust and reputation of a cloud service from consumer's point of view can be the prediction about its future behaviour based on different consumer's rating over its transaction history. Service reputation over cloud can be computed at the level of QoS parameter and service using consumer's rating. Service reputation over cloud based on QoS parameter becomes increasingly challenging when complex and mission critical applications are built upon using cloud services with diverse QoS parameters (Malik & Bouguettaya, 2009). A few QoS parameters, namely, response time changes dynamically; thereby, describing QoS of services accurately becomes complicated. A popularity based service search method is adopted, which consists of tracking all kind of services. Significant information related to discovered, invoked or published services is maintained into database to find frequency of used cloud services, their availability, usage history etc (Li, Zhou, & Yang, 2012).

A reputation as a service system is developed, in which the calculation of the reputation is no longer depending on personal experience but rather on the real-time interaction from the cloud itself (Itani, Ghali, Kayssi, & Chehab, 2014). Malicious feedback detection based on cumulative sum control charts is proposed which reduces the effect of the subjective feedback. A bootstrapping model for reputation

prediction is introduced that assigns reputation to newly deployed services in the absence of their historical information (Rathore & Suman, 2013a; 2013b; 2013c). It introduces the concept of community in bootstrapping in which cloud services related to a specific area of interest are grouped together. Cloud services of specific domain may assist each other in assigning the initial reputation of a newcomer. It uses two reputation bootstrapping approaches; namely, super-peer topology and peer to peer manner. A cumulative sum technique to identify malicious user's ratings is used and it also uses person correlation coefficient to improve the accuracy of ratings (Le-Hung, & Aberer, 2009). A trust reputation architecture is presented for the cloud computing environment to assist the decision making processes of cloud computing customers (Bilecki, L. F., & Fiorese, A., 2017). The reputed trust is based on two trust indicators: objective and subjective ones. The objective trust indicator is based on the historical QoS indicators of a cloud provider. Then, the subjective trust indicator is composed of the users' feedbacks to the cloud providers. Their future work is to apply the presented architecture in a real scenario, through a platform that offers the reputation, and analyze other aspects related to the architecture's performance, such as availability of platform, accuracy of reputation results, nodes reallocation, and analysis of other attack types.

A framework to assist reputation based discovery and matchmaking of cloud services using semantic cloud technologies is proposed (Trang, Zhao, & Yang, 2010). It consists of a communication manager, a matchmaker, a composer, and a reputation manager. In this framework, reputation manager computes the service reputation based on users' feedback and returns the service with highest reputation score based on the user's requirements. Trust and reputation model that uses Bayesian network for cloud services is presented to combine diverse sources, i.e., user's rating, QoS monitoring and direct experience opinion of the requester (Malik & Bouguettaya, 2009). It computes rater's credibility and works well even if some users are malicious. Based on historical user evaluations and their preferences related to specific requirements, a method that revises the reputation rating is proposed (Qingtao et al., 2014). Combining the prior knowledge with a filtering algorithm based on the similarities of evaluations is able to distinguish between unfair ratings in useful manner. This algorithm also includes a market mechanism that allows users and services to act as buyers and sellers. Their method uses only the user feedback ratings data related to a service.

An approach is proposed to compute cloud service reputation according to QoS history and user preferences over QoS attributes (Wang et al., 2011). The quality of rating depends on user's credibility; therefore it uses several techniques to accurately assess it. A method is stated for dynamic computation of reputation of cloud services based on their association among themselves (Elfirdoussi, Jarir, & Quafafou, 2012). The association between any two cloud services is computed based on how often they are composed together. This approach eliminates the dependency on user's feedback, but consists of overhead to maintain the dynamics of the composition and the association coefficient among cloud services. A contribution-based method is presented for propagating reputation from a composite service to its constituent services (Nepal, Malik, & Bouguettaya, 2009). Predefined weights are assigned to component services at the start of composition process. A Novel Active Cloud Service Recommendation (AWSR) approach is proposed which is based on both user's interests and preferences according to usage history to recommend cloud services actively to users (Kang, Liu, Tang, Liu, Cao, & Xu, 2012). An incentive based model is proposed that deals with reputation assessment for cloud service communities (Khosravi-far, Bentahar, Thiran, Moazin, & Guiot, 2009). A Reputation Aware QoS Prediction (RAP) approach of cloud services is presented to addresses the data credibility problem (Qiu, Zheng, Wang, Yang, & Lyu, 2013). RAP firstly calculates QoS values based user reputation contributed by service users' and then selects trustworthy values according to user reputation for prediction.

A reputation based algorithm is presented for evaluating trust which considers the feedback from the cloud service users, queue load and rejection rates of the server in detail (Challagidad, P., Reshmi, V. S., & Birje, M., 2017). Experimental results show that the presented algorithm is efficient for building trust between customers and service providers. Trust reputation architecture is presented for the cloud computing environment in order to assist the decision making processes of cloud computing consumers (Bilecki, L.F., & Fiorese, A., 2017). The presented architecture is responsible to generate aggregated reputation value, provide a module that realizes the QoS indicators monitoring, receive feedback from cloud users to providers in order to update the cloud provider's reputation and assist the decision-making processes in the cloud environment. Their future work is to apply this architecture in a real scenario, through a platform that offers the reputation. It also analyze other aspects related to the architecture's performance, such as availability of platform, accuracy of reputation results, nodes reallocation, and analysis of other attack types. A novel techniques is presented which help in detecting reputation based attacks and allowing users to effectively identify trustworthy cloud services to protect cloud services against their malevolent users and Guaranteeing the availability of the trust management service (Noor, T. H., Yao, L., Sheng, Q. Z., Anne, H.H., & Dustdar, S., 2016). Their future work plans to combine different trust management techniques such as reputation and recommendation to increase the trust results accuracy. Performance optimization of the trust management service is another focus of our future research work.

FOCUS OF THE CHAPTER

However, the trouble of trusting cloud computing is a uppermost anxiety for most organizations in such a way that trust is widely regarded as one of the top obstacles for the adoption and growth of cloud computing. Therefore, there is a need of a trust model which assists cloud consumers to find the provider that best satisfies their trust concerns in cloud computing by measuring the reliability of Cloud Service Provider (CSP). There are a variety of research work presented by existing researchers to assess and predict the service reputation in order to improve the accuracy of reputation result (Bilecki, L.F., & Fiorese, A., 2017; Noor, T. H et al., 2016; Wu, Qingtao et al., 2014; Qiu, Zheng, Wang, Yang, & Lyu, 2013). Their research mainly focuses on two challenges; firstly, computation of cloud service reputation at QoS parameter level requires that each QoS parameters should be monitored and tracked dynamically due to its dynamic nature. Secondly, dependency on cloud service user's feedback has some practical constraints, such as, complicated to guarantee the accuracy and reliability of user's feedback due to different abilities and intentions of users. Thirdly, difficult to evaluate service reputation in absence of user's feedback (Wang et al., 2011; Qiu, Zheng, Wang, Yang, & Lyu, 2013; Noor, T. H. et al., 2016; Bilecki, L.F., & Fiorese, A., 2017). In this book chapter, we have also dealt with said challenges and propose a collaborative ranking approach along with a cloud service reputation evaluation model which assesses the service reputation by collecting and combining the reputation data from multiple sources, such as access rate (Rathore & Suman, 2011, 2013, 2016, 2017) and user's rating to improve reputation result accuracy. This combination aims at reducing the probable unfairness of user ratings and helpful in improving the trustworthiness in predicting service reputation in cloud environment. Extensive experiments showed that proposed method distinguished and filtered unfair ratings correctly, while it could also recommend appropriate services for a specific user based on their preferences in a dynamic market environment.

PROPOSED COLLABORATIVE RANKING APPROACH

Improved trustworthiness of cloud service reputation does not necessarily ensured by relying on the cloud service user's feedback completely. Numerous papers have illustrated the need to use various sources of evidence due to the inherent unreliability of user feedback (Khosravifar et al., 2009; Wang, Sun, Zou, & Yang, 2011). Reliability in prediction of service reputation can be obtained by considering one more opinion about service reputation and combine it with the reputation obtained through user's feedback ratings. In other words, one opinion can be obtained by performing active monitoring of services which are accessed by service consumer. The second opinion is dependent on user's rating which are provided by service users' manually about his/her direct experience with the service after the successful invocation. Figure 1 shows the cloud service reputation evaluation model which shows the overall process of proposed approach. The presented model is applicable to most of the ratings-based experiences shared on online platforms where users evaluate services with numerical ratings.

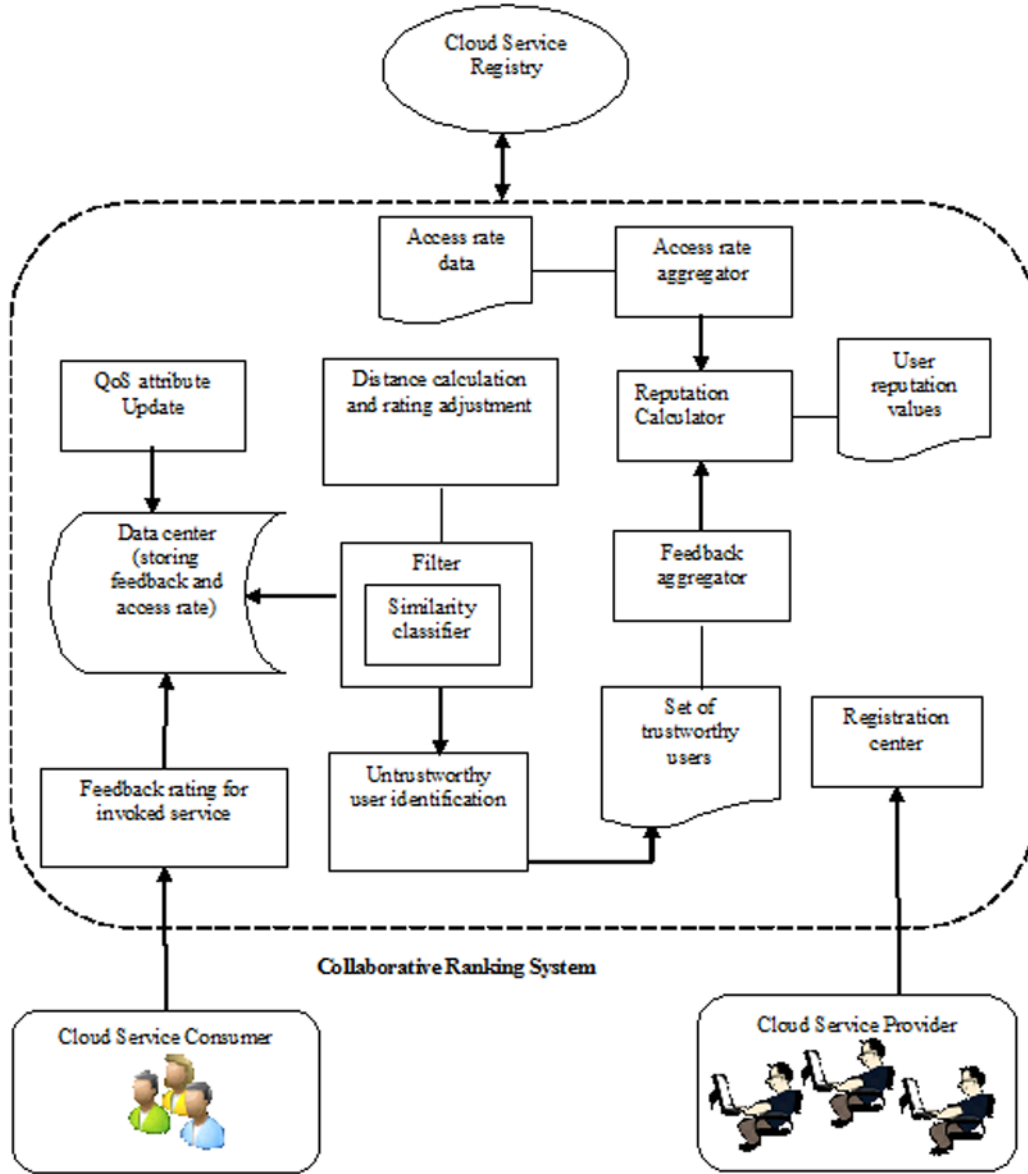
The cloud service providers and the cloud service consumers make up the cloud environment with the collaborative ranking system which playing the central role. Trust management comprises three roles, such as the Data Center, Registration Center and Reputation Calculator. The cloud Service providers and the cloud service consumers are all required to register in the registration center before they enter the cloud marketplace. Registration requires provision of essential information which is stored in the data center. For cloud service providers, the essential information includes the cloud service's ID, the function that the service provides and the offered performance of the function. For cloud consumers, it requires personal information typical for user profiles, which we assume is confidential in our model. The Data Center stores the essential information with access rate and also records feedback given by cloud service consumers who experienced a service.

Feedback is denoted as a tuple $F = (Sid, Uid, R, T)$, where Sid represents the identity of the cloud service provider whose services are invoked by the consumer and Uid is the identity of the cloud service consumer who used the Sid. Feedback rating is represented by R provided by the service consumer with integer values in the range (Xu et al., 2015; Elfirdoussi, Jarir & Quafafou, 2012; Li, Zhou, & Yang, 2012; Li, & Wang, 2011). T is a timestamp used to records the time when the feedback was given by the consumer.

The key component in proposed approach is Reputation Calculator, which is able to aggregate the feedback for a cloud service into a reputation score for the service requestor. It is based on the resemblance of preferences between the cloud service requestor and feedback givers. In general, it is unstated that users with higher similarity in their attributes would have a higher probability for attainment of agreement on an indistinguishable thing. Hence, the consumers' with similar feedbacks are more reliable than the universal view in cloud environment.

In this figure, user ratings are collected as feedback to reputation processing node. Following the normalization of feedback data, the feedback ratings are filtered on the basis of similarity categorization. Next, the preferences for abnormal cloud users are set for their specific service requirements. The user preferences mean that the recommendation system provides the most relevant services. There is also a convinced extent of punishment for collusive users, who makes the outcome puzzling when service recommendations are required. The core of the system is the reputation calculation, where we combine the result from filtering with the historical reputation to generate a reputation value that is credible and reliable.

Figure 1. Cloud Service Reputation Evaluation Model



The presented model evaluates the reputation of a service as follows:

$$R(S) = \frac{\sum_{i=1}^n f(i, s) * R_i}{n} \quad (1)$$

Where, $R(s)$ represents reputation score of cloud service s and $f(i, s)$ is the feedback to the s^{th} cloud service from the i^{th} cloud consumer. R_i is the rating of the i^{th} cloud consumer, which is computed based on the similarity between the requestor for the s^{th} cloud service and consumers who have experienced

the s^{th} cloud service and given the feedback. N is the number of consumers who rated the particular cloud service.

Assessing Service Reputation Using Access Rate

Let, $R_{(w_j)}$ is the reputation of a cloud service w_j at time instance t within the limit 0 and 1, i.e., $\forall R_{(w_j)}(t) \in (0,1) \forall R_{l(w_j)}(t) \in (0,1)$. Also, let $R_{(w_j)}(t-1)$, $1 \leq j \leq n$ be the reputation level of w_j at time instance $t-1$, i.e., $\forall R_{(w_j)}(t-1) \in (0,1)$. Let, A_r be the access rate obtained through active monitoring of service invocations w_j to m number of times at time instance t , i.e.,

$$\forall A_r(w_j) \in (1, m) \quad (2)$$

$$\forall R_{l(w_j)}(t-1) \in [0,1]$$

In order to update the old service reputation and compute the newer one $R_{(w_j)}^i(t)$ of a cloud service w_j based on its access rate at t can be obtained as,

$$R_{(w_j)}^i(t) = f(R_{(w_j)}(t-1), A_r, \cup, U_r) \quad (3)$$

$$f(A_r) = R_{(w_j)}^i(t) = \frac{\sum_{i=1}^m f(i, s) * A_r}{m} \quad (4)$$

Where, $f(i, s)$ is the feedback to the s^{th} cloud service from the i^{th} cloud consumer. $R_{(w_j)}(t-1)$ is a previous reputation value at time instance $t-1$, U_r is the user feedback rating, $\cup$ is the propositional logic operator and $R_{(w_j)}^i(t)$ is the i^{th} reputation level at t .

Detection of Malicious User Feedback Rating Using Similarity Computation

Reputation of a cloud service can be determined according to feedback ratings provided by service users. A collaborative filtering method uses user feedback in the evaluation of reputation accuracy of cloud services. Therefore, feedback ratings given by service consumers can not be entirely avoided. But the major issue with user feedback rating is the trustworthiness and accuracy of their feedback. One possible solution towards this problem is to provide an automatic mechanism for detection of malicious user feedback to assess the service reputation by discarding it.

The user feedback rating data is separated into two parts such as, malicious feedback ratings and actual feedback ratings. It can be achieved by setting up the maximum and minimum threshold value for malicious and normal feedback ratings. If the rating provided by majority of users is negative feedback rating then the service reputation is computed on the basis of negative feedback ratings and considered to be of poor reputation. If majority of consumers provides genuine feedback ratings then the service

reputation is computed based on genuine feedback ratings. Here, we have used the simple averaging method to compute the service reputation based on actual user feedback ratings. If the service reputation is evaluated using genuine feedback ratings then this is compared against the ideal reputation in order to include a service in service composition. In this phase, feedback ratings are considered only from trustworthy users for reputation assessment.

Here, manual user feedback is collected from group of consumers for their usage experience with the service. A log file is maintained for user feedback data, with columns user_id and user_feedback rating. Feedback rating is any numeric value between 1 to 5, in which 1 means poor service, 2 means average, 3 means good, 4 means very good and 5 means excellent. Here, feedback data from 15 consumers are collected and analyzed to determine whether it contains any malicious user rating or not. It is necessary to consider only that users rating, which are accurately provided in order to improve the accuracy of reputation of a service.

In online dating service the user feedback ratings are analyzed to identify the occurrence of positive or negative malicious rating providers. For example, if a service user provides the feedback of a service as excellent and majority of users provide the ratings as poor, then the malicious positive rating is removed for further computation of reputation of a service to predict and improve reputation accuracy. This phase works in two steps, first step identifies the list of non malicious users by performing similarity computation of previous user's feedback and second step compares the percentage of non malicious users with the specified threshold value. Related feedback ratings are recognized with similarity computation of previous user's feedback using Pearson correlation coefficient (PCC). Identical feedback ratings are recorded in similarity coefficient variable. Consider that there are m users and n cloud services, and the relationship between users and cloud services is denoted with an m*n matrix. Then each matrix entry $r_{x,i}$ denotes the user feedback rating of the cloud service I rated by the user x, where, $r_{x,i}$ is a normal feedback rating. Below equation is used to calculate the similarity among user x and user y on the basis of their frequently rated cloud services:

$$\text{Sim}(x,y) = \sum_{i \in I_x \cap I_y} (r_{x,i} - \bar{r}_x)(r_{y,i} - \bar{r}_y) / \sqrt{\sum_{i \in I_x \cap I_y} (r_{x,i} - \bar{r}_x)^2} * \sqrt{\sum_{i \in I_x \cap I_y} (r_{y,i} - \bar{r}_y)^2} \quad (5)$$

Where, $\text{Sim}(x,y) \in [-1,1]$ represents the similarity of two users. $I_x \cap I_y$ is a set of commonly rated cloud services by both users a and u. $r_{x,i}$ and $r_{y,i}$ are the two feedback ratings of cloud service i rated by user a and u. $\bar{r}_x$ represents the average feedback rating of all of the cloud services that are rated by user x.

Calculation and ranking of PCC similarity values stuck between the current user and the other users, a set of similar users $S(x)$ can be identified, as follows:

$$s(x) = \{y | \text{sim}(x,y) \geq \text{sim}_k, \text{sim}(x,y) > 0, x \neq y\} \quad (6)$$

Where, Sim_k is the k^{th} largest PCC value with the current user y. k represents the number of the similar users (i.e., higher PCC values than others), and $\text{sim}(x,y) > 0$ is to eliminate the unlike users (dissimilar users, e.g., those with negative PCC value, will influence the reputation measurement accuracy).

After obtaining the set of similar users, according to a set of community cloud services $SS_k = s1k, s2k, \dots, slk$ which contains one services used by the K number of users, the feedback similarity between user a and user u can be calculated as:

$$f(U_r)=FSk(x,y)=\begin{cases} 1-\sqrt{\frac{\sum_{yTS(x)}(r_{x,i}^k-r_{y,i}^k)}{|SS^k|}} & \text{if } |SS^k| \neq 0. \\ 0 & , \text{if } |SS^k| = 0 \end{cases} \quad (7)$$

$$f(U_r)=\frac{\sum_{i=1}^n f(x,i)*R_i}{n} \quad (8)$$

Collaborative Service Reputation Using Propositional Logic

The reputation values obtained from two different sources can further be combined in order to determine final service reputation. Here, $f(U_r)$ represents a function in which service reputation is computed through user feedback and $f(A_r)$ represents a function in which service reputation is computed through runtime access rate. Therefore, the average of the values returned from both the functions $f(A_r)$ and $f(U_r)$ can be taken, and the composition of these two functions generate the final service reputation . Only the services with above a given maximum threshold value ($\text{Max}_{\text{threshold}}$) are provided to cloud service consumer; otherwise, below the threshold value, service is not trustable and not able to discover. In both the cases, old service reputation is updated by the newer one. The reputation of a service can be obtained as:

$$R = [f(U_r) + f(A_r)]/2$$

[Update old service reputation].

$$R_{(w_j)}^i(t) = f(R_{(w_j)}(t-1), A_r, U_r)$$

```

If ( $R \geq \text{Max}_{\text{threshold}}$ ) then
  Discover_service();
Else
  Remove_service (webservice);
End if

```

EXPERIMENTAL ANALYSIS

For performing reliable experiments, we have used large-scale real cloud service dataset named WS-DREAM as shown in Table 1 (Wang et al., 2011; Rathore & Suman, 2013). WS-DREAM dataset is termed as Distributed Reliability Assessment Mechanism which allows cloud service users to carry out service reliability and quality assessment in a collaborative manner. This dataset maintains the service invocation records of 1.5 millions services of 150 users in 24 countries. It also collects non-functional

attributes (e.g., QoS) of these cloud services observed by 339 distributed computers positioned in 30 diverse countries, from Planet-Lab. Here, the number of service invocation parameter of WS-DREAM dataset is considered as access rate. User's rating for some cloud services are directly collected from seekda.com, service-repository.com and xmethod.com. For those services whose rating scores are not available from these sites, we determined their rating scores (e.g., from 1 to 5) based on their OAE-QoS values score (Yuan, An, & Wang, 2009; Rathore & Suman, 2013, 2016, 2017). Here, 1 means poor service, 2 means average, 3 means good, 4 means very good and 5 means excellent.

A list of experiments has been conducted on the proposed ranking mechanism to assess and predict the service reputation over cloud. The performance of presented mechanism is compared with the algorithms proposed in existing approaches (Trang, Zhao, & Yang, 2010; Le-Hung, & Aberer, 2009; Wang, Sun, Zou, & Yang, 2011; Brozovsky, & Petricek, 2007). Mean Absolute (MAE) and Root Mean Square (RMSE) Error is employed as measurement standard to predict service reputation of diverse approaches (Wang, Sun, Zou, & Yang, 2011). MAE and RMSE can be helpful in determining the deviation between the predicted and the observed error in cloud services reputation. The presented approach assess the service reputation based on both manual user's feedback rating and runtime service access rate parameters of WSDREAM dataset. Two experiments are performed in this section to assess the performance of proposed ranking approach. The first experiment is performed on reputation accuracy, whereas the second experiment is performed on deviation of each service from ideal reputation.

Table 1. WS-DREAM QoS dataset

SMS			Captcha			Send mail WSDL			Calculator			Buyer Data			Echo		
user id	No. of invocation	User rating	user id	No. of invocation	user rating	user id	No. of invocation	user rating	user id	No. of invocation	user rating	user id	No. of invocation	user rating	user id	No. of invocation	user rating
12.108.127.136	676	3	35.9.27.26	642	5	35.9.27.26	969	3	64.151.112.20	791	5	72.36.112.71	837	5	128.2.223.63	632	5
12.108.127.137	587	4	35.9.27.27	642	5	35.9.27.27	605	4	64.151.112.21	817	5	72.36.112.72	623	5	128.2.223.64	595	5
12.108.127.138	678	3	35.9.27.28	587	5	35.9.27.28	631	3	64.151.112.22	649	5	72.36.112.73	601	5	128.2.223.65	5911	5
12.108.127.139	564	1	35.9.27.29	928	5	35.9.27.29	638	1	64.151.112.23	711	4	72.36.112.74	1439	5	128.2.223.66	9273	5
12.108.127.140	1408	2	35.9.27.30	561	5	35.9.27.30	623	2	64.151.112.24	616	4	72.36.112.75	1427	5	128.2.223.67	10134	5
12.108.127.141	1471	2	35.9.27.31	587	5	35.9.27.31	601	2	64.151.112.25	605	4	72.36.112.76	1499	5	128.2.223.68	713	5
12.108.127.142	1436	1	35.9.27.32	2133	5	35.9.27.32	10134	1	64.151.112.26	631	4	72.36.112.77	1508	5	128.2.223.69	614	5
12.108.127.143	1428	4	35.9.27.33	594	1	35.9.27.33	713	4	64.151.112.27	2096	4	72.36.112.78	712	1	128.2.223.70	454	1
12.108.127.144	969	5	35.9.27.34	738	4	35.9.27.34	614	5	64.151.112.28	599	4	72.36.112.79	1473	4	128.2.223.71	631	4
12.108.127.145	605	5	35.9.27.35	630	4	35.9.27.35	1368	5	64.151.112.29	645	4	72.36.112.80	930	4	128.2.223.72	595	4
12.108.127.146	631	5	35.9.27.36	759	4	35.9.27.36	601	5	64.151.112.30	25372	4	72.36.112.81	703	4	128.2.223.73	728	4
12.108.127.147	638	5	35.9.27.37	556	1	35.9.27.37	816	5	64.151.112.31	3331	4	72.36.112.82	923	5	128.2.223.74	2693	5
12.108.127.148	1408	4	35.9.27.38	1874	5	35.9.27.38	1396	4	64.151.112.32	587	4	72.36.112.83	454	5	128.2.223.75	620	5
12.108.127.149	1471	4	35.9.27.39	716	5	35.9.27.39	674	4	64.151.112.33	1368	5	72.36.112.84	631	4	128.2.223.76	1041	4
12.108.127.150	1436	3	35.9.27.40	700	5	35.9.27.40	623	3	64.151.112.34	601	5	72.36.112.85	5911	4	128.2.223.77	928	4
12.108.127.151	1428	3	35.9.27.41	651	5	35.9.27.41	616	3	64.151.112.35	816	5	72.36.112.86	9273	4	128.2.223.78	561	4
12.108.127.152	969	3	35.9.27.42	38374	5	35.9.27.42	812	3	64.151.112.36	1396	5	72.36.112.87	10134	4	128.2.223.79	587	4
12.108.127.153	605	3	35.9.27.43	600	5	35.9.27.43	832	3	64.151.112.37	674	5	72.36.112.88	713	4	128.2.223.80	2156	4
12.108.127.154	631	3	35.9.27.44	1416	3	35.9.27.44	1400	3	64.151.112.38	712	5	72.36.112.89	614	4	128.2.223.81	594	4
12.108.127.155	638	3	35.9.27.45	824	4	35.9.27.45	707	3	64.151.112.39	1473	5	72.36.112.90	605	4	128.2.223.82	738	4
12.108.127.156	928	2	35.9.27.46	688	3	35.9.27.46	25372	5	64.151.112.40	930	1	72.36.112.91	631	4	128.2.223.83	630	4

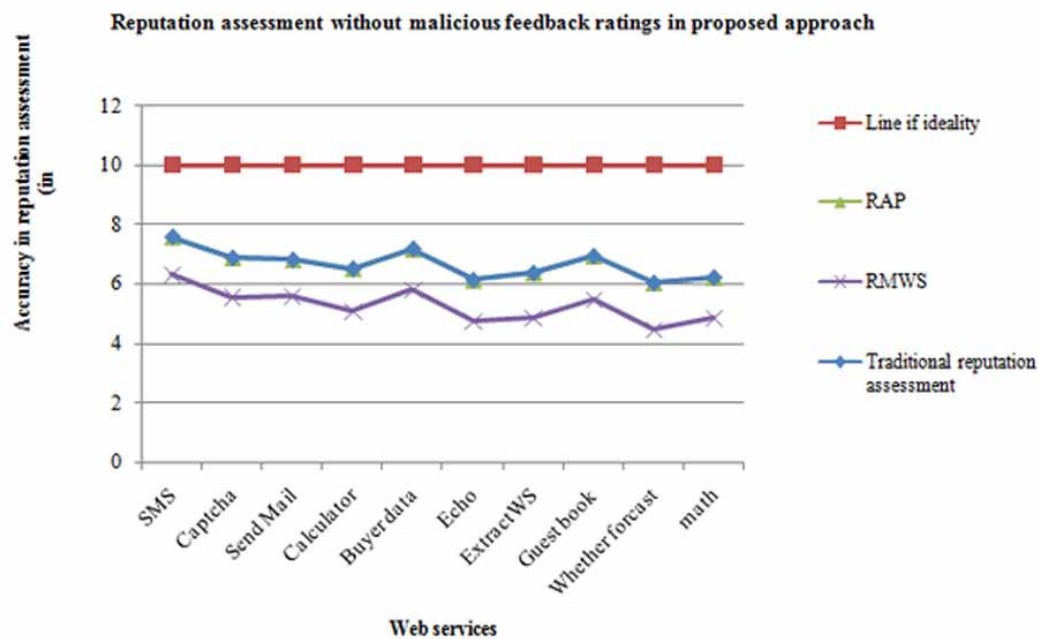
Experiment on Accuracy

In this experiment, existing approaches that uses user's feedback rating for reputation assessment of cloud services are compared with the proposed one in order to determine the improved service reputation over cloud (Trang, Zhao, & Yang, 2010; Le-Hung, & Aberer, 2009; Wang, Sun, Zou, & Yang, 2011; Brozovsky, & Petricek, 2007). Existing approaches use user's feedback rating for the service reputation assessment without excluding the malicious feedback ratings, whereas proposed approach detect malicious feedback ratings and computes the cloud service reputation by discarding it. The performance of

presented approach in assessing the service reputation is better than the existing approaches as it does not consider malicious ratings for reputation assessment.

In Figure 2 and Figure 3, ideality line represents the ideal reputation through which the performance of RAP, Reputation Measure approach of Cloud Services (RMWS) and proposed approach can be effectively evaluated. Figure 2 shows the results of service reputation assessment in which only user's feedback rating are included. In Figure 2, there is a significant improvement of approximately 10% in reputation prediction in traditional and existing approaches when considering all feedback ratings (malicious and actual). In traditional reputation assessment mechanism and RAP, service reputation is equal i.e., less than 80%, whereas, in RMWS, it is below 70%. Figure 3 shows the results of assessing cloud service reputation in which the malicious ratings are discarded to provide improved service reputation. As compared to Figure 2, Figure 3 shows the improved reputation accuracy in proposed ranking approach by 20% from RMWS and 10% from RAP. In both the figures, line of ideality represents ideal reputation which is the maximum allowable reputation level.

Figure 2. Accuracy in service reputation in proposed and existing approaches using all user feedback rating



Practically, it is not possible to determine the ideal reputation; instead it can be possible to consider the maximum threshold value to be followed. This means that smaller the deviation is, the approach is more close to the ideality line. As shown in Figure 4, the deviation in proposed approach is smaller than the existing approaches from ideality line. In Figure 4, the deviation in proposed approach is 2.0455 on average, whereas, the deviation of RMWS and RAP on average are 4.737, 9.333 respectively. We compared the performance of proposed approach with the existing approaches under MAE and RMSE measurement criteria (Wang, Sun, Zou, & Yang, 2011). As the dataset is too large to perform experiment, therefore we have considered only a limited number of services which are selected on a random

Figure 3. Accuracy in service reputation assessment in proposed and existing approaches using user feedback rating

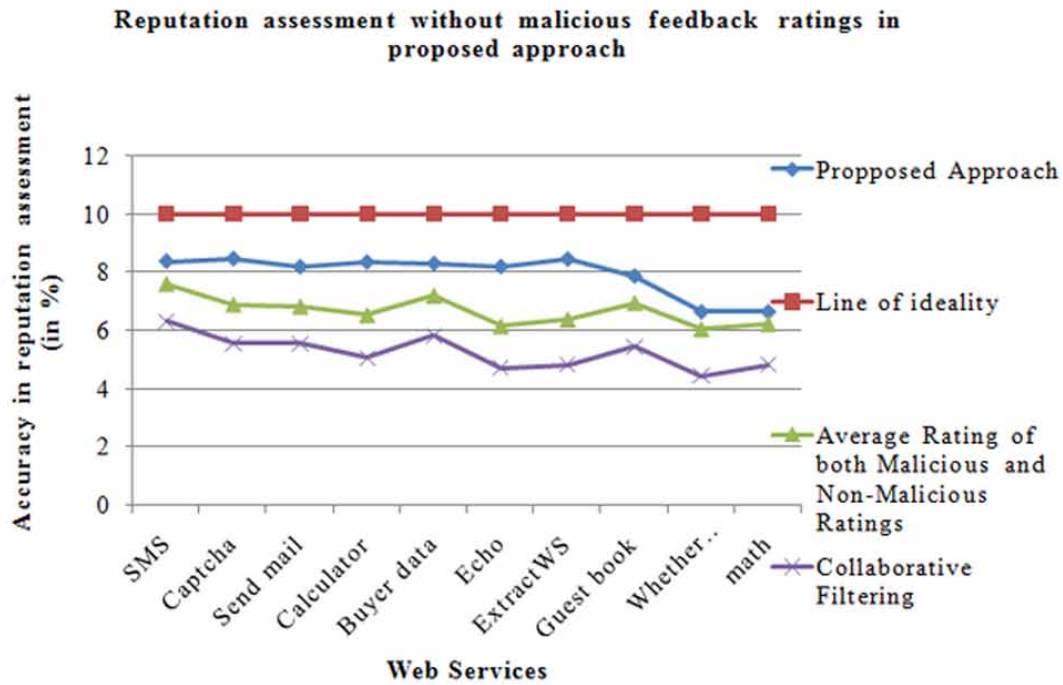


Figure 4. Reputation prediction accuracy using MAE

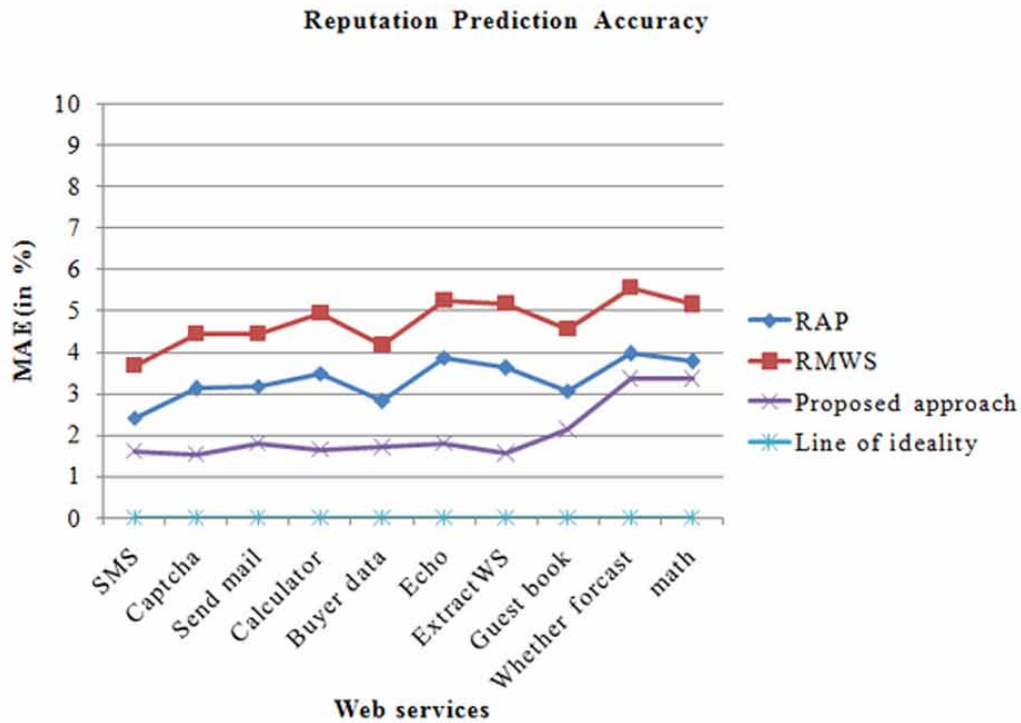
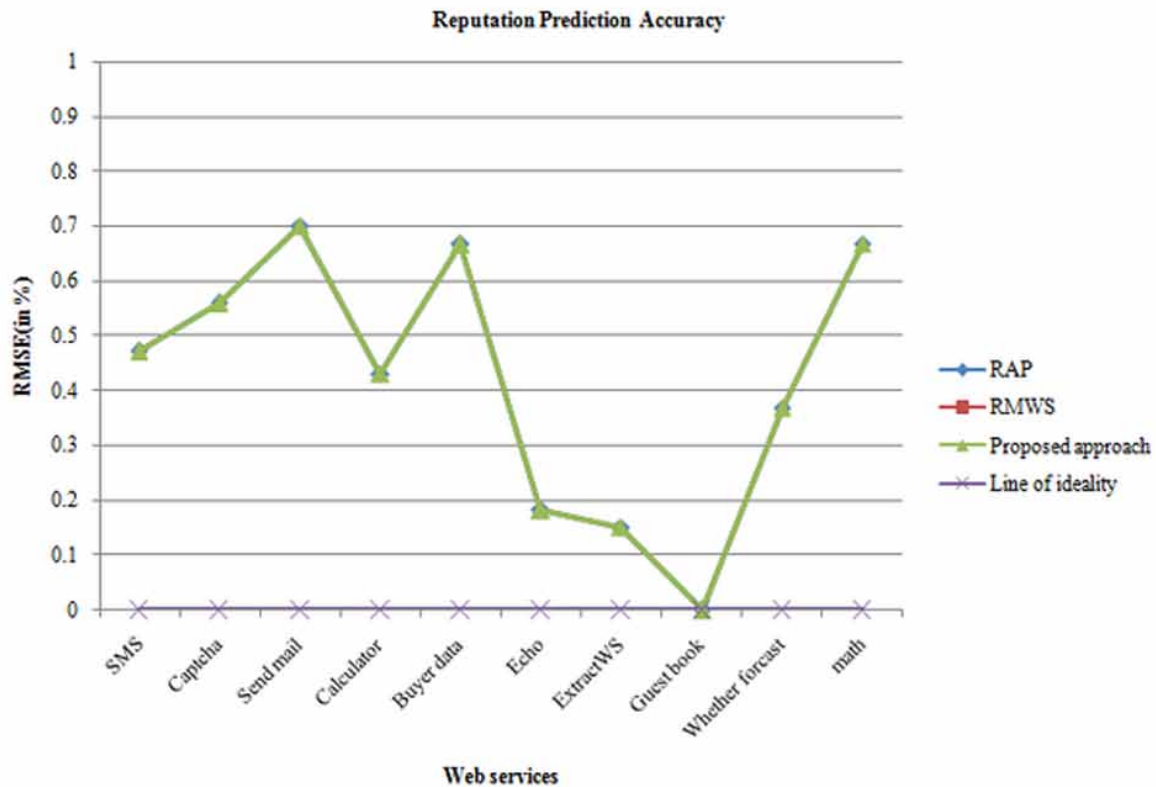


Figure 5. Reputation prediction accuracy using RMSE



basis from the WSDREAM dataset. It can be observed that the MAE and RMSE are smaller in proposed ranking mechanism as compared to RAP and RMWS. A smaller value of MAE or RMSE indicates better service reputation prediction. Figure 4 and Figure 5 shows that the effect of MAE and RMSE is less in proposed approach whereas, RAP and RMWS both has high and equal effect.

FUTURE RESEARCH DIRECTIONS

A series of experiments have been performed to validate the efficacy of proposed approach. It is compared with the existing approaches on the basis of MAE and RMSE measurement criteria in order to determine the deviation between ideal and predicted service reputation. It is observed that the proposed approach has smaller MAE and RMSE as compared to existing approaches from the line of ideality. In practice, it is difficult to determine the ideal reputation, instead a threshold value can be considered to evaluate and validate the efficiency of proposed approach. Proposed ranking mechanism ensures trustworthiness in reputation prediction by combining two different sources to determine the common decision on reputation of service. Although the proposed approach is effective than existing approaches, it also has some limitations, such as overhead of maintaining the records from multiple sources will be increased. There are a few directions for our future work involves combining different trust management techniques to increase the accuracy of trust results and decreasing the overhead of maintaining the records from multiple sources.

CONCLUSION

With the ever increasing services available over cloud, it becomes very difficult to choose the reliable services as per the cloud service consumer's requirements. In this book chapter, a collaborative ranking approach along with a cloud service reputation evaluation model is proposed for reliable discovery and selection of suitable cloud services. It assesses the service reputation using two equally significant parameters, such as access rate and user's rating. It is helpful in reducing the probable unfairness of user ratings and improves the trustworthiness of predicting service reputation in cloud environment. It is also able to detect malicious user's rating and compute service reputation by including only actual user ratings. Proposed approach is validated by conducting several experimental studies using 3,693 real-world services publicly available from the internet. It shows that presented approach distinguished and filtered unfair ratings correctly, as well as it could also recommend suitable services to a specific user based on their preferences in a dynamic market environment.

REFERENCES

- Bilecki, L. F., & Fiorese, A. (2017). A Trust Reputation Architecture for Cloud Computing Environment. *Proceedings of 14th International Conference on Computer Systems and Applications (AICCSA)*, 614-621. 10.1109/AICCSA.2017.31
- Brozovsky, L., & Petricek, V. (2007). Recommender system for online dating services. *Proceedings of 6th conference (Zalosti' 07)*.
- Challagidad, P., Reshmi, V. S., & Birje, M. (2017). Reputation Based Trust Model in Cloud Computing. *Advances in Cloud and Internet of Things*, 5(5), 5-12.
- Elfirdoussi, S., Jarir, Z., & Quafafou, M. (2012). Popularity Based Web Service Search. *Proceedings of 19th IEEE International Conference on Web Services (ICWS)*, 683-689.
- Itani, W., Ghali, C., Kayssi, A., & Chehab, A. (2014). *Reputation as a Service: A System for Ranking Service Providers in Cloud Systems Security, Privacy and Trust in Cloud Systems*. Springer Berlin Heidelberg.
- Kang, G., Liu, J., Tang, M., Liu, X., Cao, B., & Xu, Y. (2012). AWSR: Active Web Service Recommendation Based on Usage History. *Proceedings of 19th IEEE International Conference on Web Services (ICWS' 12)*, 186-193.
- Khosravifar, B., Bentahar, J., Thiran, P., Moazin, A., & Guiot, A. (2009). An Approach to Incentive-Based Reputation for Communities of Web Services. *Proceedings of IEEE International Conference on Web Services (ICWS' 09)*, 303-310. 10.1109/ICWS.2009.101
- Le-Hung, V., & Aberer, K. (2009). Towards Probabilistic Estimation of Quality of Online Services. *Proceedings of IEEE International Conference on Web Services (ICWS' 09)*, 99-106.
- Li, L., & Wang, Y. (2011). A subjective probability based deductive approach to global trust evaluation in composite services. *Proc. of the International Conference on Web Services (ICWS)*, 604-611. 10.1109/ICWS.2011.28

- Li, X., Zhou, F., & Yang, X. (2012). Scalable feedback aggregating (SFA) overlay for large-scale P2P trust management. *IEEE Transactions on Parallel and Distributed Systems*, 23(10), 1944–1957. doi:10.1109/TPDS.2012.26
- Malik, Z., & Bouguettaya, A. (2009a). RATEWeb: Reputation Assessment for Trust Establishment among Web Services. *The VLDB Journal*, 18(4), 85–911. doi:10.1007/00778-009-0138-1
- Malik, Z., & Bouguettaya, A. (2009b). Reputation Bootstrapping for Trust Establishment among Web Services. *Proceedings of IEEE Internet Computing Conference, IEEE Computers & Society*, 13(1), 40–47.
- Nepal, S., Malik, Z., & Bouguettaya, A. (2009). Reputation propagation in composite services. In *Proceedings of 7th IEEE International Conference on Web Services*. IEEE Computer Society.
- Noor, T. H., Yao, L., Sheng, Q. Z., Anne, H. H., & Dustdar, S. (2016). CloudArmor: Supporting Reputation-Based Trust Management for Cloud Services. *IEEE Transactions on Parallel and Distributed Systems*, 27(2), 367–380. doi:10.1109/TPDS.2015.2408613
- Qiu, W., Zheng, Z., Wang, X., Yang, X., & Lyu, M. R. (2013). Reputation aware QoS value prediction of Web services. In *Proceedings of IEEE International Conference on Services Computing (SCC' 13)*. IEEE Computer Society. 10.1109/SCC.2013.43
- Rathore, M., & Suman, U. (2011). A quality of service broker based process model for dynamic web service composition. *Journal of Computational Science*, 7(8), 1267–1274. doi:10.3844/jcssp.2011.1267.1274
- Rathore, M., & Suman, U. (2013a). An ARSM approach using PCB-QoS classification for web services: A multi-perspective view. In *Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE Computer Society. 10.1109/ICACCI.2013.6637165
- Rathore, M., & Suman, U. (2013b). Web service selection algorithm for dynamic service composition using LSLO approach. In *Proceedings of International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE Computer Society. 10.1109/ICIEV.2013.6572688
- Rathore, M., & Suman, U. (2013c). Evaluating QoS parameters for ranking Web service. In *Proceedings of 3rd International Advance Computing Conference*. IEEE Computer Society. 10.1109/IADCC.2013.6514438
- Rathore, M., & Suman, U. (2016). A Mixed Opinion based Reputation Prediction Approach for Web Service Discovery. *International Journal of Business Information Systems*, 22(2), 143–165.
- Rathore, M., & Suman, U. (2017). Classification of Non-Functional Requirements of Web Services from Multi-Perspective View. In *RESCC Book 2017: CfC on Requirements Engineering for Service and Cloud Computing* (pp. 65–87). Springer International Publishing.
- Trang, N. H., Zhao, W., & Yang, J. (2010). A Trust and Reputation Model Based on Bayesian Network for Web Services. *Proceedings of IEEE International Conference on Web Services*, 251–258.
- Wang, P., Chao, K-M., Lo, C-C., Lin, W-H., & Wang, T. (2011). A Consumer Trust-Based Approach in Web Services Selection. *Proceedings of 8th IEEE International Conference on e-Business Engineering*, 251–256. 10.1109/ICEBE.2011.15
- Wang, S., Sun, Q., Zou, H., & Yang, F. (2011a). Reputation measure approach of Web service for service selection. *IET software (Beijing, China)*, 5(5), 466–473.

Wang, S., Zheng, Z., Sun, Q., Zou, H., & Yang, F. (2011b). Evaluating Feedback Ratings for Measuring Reputation of Web Services. *Proceedings of IEEE International Conference on Services Computing*, 192-199. 10.1109/SCC.2011.32

Wu, Q. (2014). Reputation Revision Method for Selecting Cloud Services Based on Prior Knowledge and a Market Mechanism. *The Scientific World Journal*, 14, 9.

Xu, J., Jiang, D., Wang, B., Yang, D., & Reiff-Marganiec, S. (2015). Local Reputation Management in Cloud Computing. *2015 IEEE World Congress on Services*, 261-267. 10.1109/SERVICES.2015.45

Yuan, J., An, Y., & Wang, H. (2009). Reputation-Based Semantic Service Discovery. *Proceedings of International Conference on Multimedia Information Networking and Security*, 61-64.

KEY TERMS AND DEFINITIONS

Access Rate: The rate of total number of cloud service request requested by the cloud service consumer.

Cloud Computing: A platform where one can find all kind of services at one place through internet.

Cloud Service Consumer: An entity who utilizes cloud services.

Cloud Service Provider: An entity who offers cloud services over cloud.

Collaborative Filtering: An automated predictive method of collecting preferences from many users based on their interest.

OAE-QoS: An overall aggregated score of all QoS parameters of a cloud services.

Quality of Service: The overall description of the performance of a cloud service.

Rater's Credibility: The quality of the service that makes cloud service consumers or providers to trust the information or service.

Reputation: An opinion about a cloud service by cloud consumer based on its previous history.

Root Mean Square and Mean Absolute Error: The deviation between the predicted and the observed error in measuring cloud services reputation.

Trust: A belief among cloud service consumer and provider about the reliability of cloud services.

Tuning Drone Data Delivery and Analysis on the Public Cloud

2

Jose Lo Huang

Simon Bolivar University, Venezuela

INTRODUCTION

There has been great growth in the use of drones for research. The drone is “a remote-controlled pilotless aircraft or missile” capable of visualizing extended areas of study in less time than the human eye (OxfordDictionaries.com, 2018). Many drones can send live video streaming to mobile phones or computers. Data collected by drones provides useful information and patterns to deploy new technologies and generate deeper data analysis. These data, including information on movement, color, or area patterns, can be processed per frame and mapped as color pixels or represented as a set of Extensible Markup Language (XML) files.

According to Mell and Grance (2011, p. 2) from the National Institute of Standards and Technology (NIST):

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The public cloud, such as Amazon Web Services (AWS) or the Google Cloud Platform (GCP), offers reliable and nearly unlimited storage (www.aws.amazon.com). Paid on-demand services can be used as additional storage or compute services. Users can create several parallel virtual machines and process data loads through the public cloud. It is possible to automatically scale the use of virtual machines depending on resource needs (www.aws.amazon.com).

Photogrammetry is the use of photography in surveying and mapping to ascertain measurements between objects (OxfordDictionaries.com, 2018) and is an example of how a machine can be used to process images collected from drones. Monthly data can be efficiently processed by integrating and analyzing terabytes (TB) or petabytes (PB) from drones on the public cloud. However, there is not a faster way to send large amounts of data from a premise server to the public cloud.

The current method of processing data from drones on the cloud use raw data and these were sent in sequential order. This method depends on the speed of transference of every packet from the base station to the cloud. Also, if the company use only one node in the cloud for processing, then the overall process is sequential and very slow. This amount of time could be very long with a regular video of 300 MB. Using a set of videos of regular size can take several hours or days, depending on the bandwidth and the flow between the intermediate network devices.

In this article, the researcher will use a set of tools and an automation process to study a unique method of compress the output of drone videos and improve the transference of data from the on-premise base station to the cloud to analyze them in a concurrent environment on the public cloud. After discussing alternatives, the chapter will review process data calculations to compare faster methods. With modifica-

DOI: 10.4018/978-1-7998-3479-3.ch016

tions to upload times, the analyzed data time is reduced by an average of 78% as compared to a common sequential process. This valuable time can be used by companies to be focused on their experiment and business results instead of the processing of the data sets.

BACKGROUND

Several studies have discussed the integration of drones or robots and the public cloud. In the internet of drones theory (Gharibi, Boutaba, & Waslander, 2016), the authors created a system to manage and control access to drones and their activities to navigate service locations. A study by Yinong and Garcia (2010) reviewed the field of robots as a service in cloud computing. Through distributed systems, the researchers controlled a network of segregated robots and sent various programs to the robots to deliver different services.

In a service-oriented architecture (SOA) study, researchers created a Web-based system hosted in the public cloud to manage drone networks (Koubaa, Qureshi, Sriti, Javed, & Tovar, 2017). In a robust study by Mahmoud, Mohamed, and Al-Jaroodi (2015), the authors designed a technological system using unmanned aerial vehicles (UAVs) as nodes based on the theory of the internet of things (IoT). Using the RESTful HTTP technology and Arduino boards, the UAVs transmitted data with the internet.

Just a few years before these studies, pilots could not use drones from a distance. However, distance no longer limits the user. It is common practice to use data collected by sensors in wireless devices. For example, Lo, Lopez, Goncalves, and Perera (2015) fetched data with several types of sensors analyzed with the Skyline algorithm. This study had collected a large amount of data from sensors and stored them in local systems to be analyzed with the algorithm. Although, with this method the researchers had to have the hardware to process the data. Nowadays, data scientists have the option to take advantage of the better and on demand resources in the cloud to analyze faster and a huge quantity of data. For example, in a research made by Saleem, Salim, and Rehmani (2014) they explain the integration between Wireless Sensor Networks (WSN) and the public cloud. Applications of WSN over the cloud are then described in that study.

The development of theories about the improvement of the work done on the cloud was studied in different researches. For instance, the research made by Bitam (2012) where the author has been studied and solved, with the Bees Life Algorithm, the job scheduling problem in the cloud using a method simulating the behavior of the bees in the nature. Another interesting study can be found in the research about a cloud-enabled robotic system for real-time video tracking applications made by Liu, Chen, Pham, Shen, and Chen (2014). In this study, the researchers used effectively the cloud computing resources to accomplish more computationally intensive tasks of a set of robotic systems. They used a cloud enabled robotic system (CERS) framework to process images from a processing application.

There are some studies about the processing of images on the cloud with the security as the main concern. For example, in the research conducted by Sattar, Adnan, and Kali (2017), the focus of the experiment is maintaining a secure environment inside the cloud when processing a data set. In a research developed by Welsh (2016), the author compares the processing time of a big set of images in local hardware (8 weeks) and the same data set in a server farm on the cloud (few hours). He used high quality images in different formats as 3D or 4K.

Although researchers have studied image processing, drones and the public cloud before, they have not mentioned the latency and time to send data from on premise to the cloud. This may be due to small data samples or the exclusion of this information from their studies. When research includes large amounts of

data, there is a need to tune and improve transmission times between local machines and public clouds because the time spent in the transference is high. In this paper, the author presents a robust method to tune the process of send and analyze the images from drones on the public cloud.

FOCUS OF THE ARTICLE

It is common for scientific studies to use drones to review images in a specific area. After the videos was completed, they stored them in local devices to process them or take advantage of the public cloud to gain more resources to speed up the process of the final videos. When they use the public cloud, the more common way of send the information is in sequential mode. This is, one by one using a First In First Out (FIFO) mechanism. And when the videos arrived at the cloud, the common method to process them is using one server instance with enough resources in terms of CPU and RAM memory.

With large amount of data sets, this sequential process is not enough to analyze the data in a good time. For this reason, this paper develops a new way to transfer the images and processing them on the cloud. This process includes techniques as compression, concurrency and autoscaling.

In this article, the author explains with an experimental work, how to tune two core processes to analyze drone data in the public cloud. First, the study analyzes and tunes the transmission of drone data to an on-premise base station, as well as the transmission of data from an on-premise station to public cloud machines. Second, tunes the process of analyzing data on the public cloud. In the end of the research, there are some results about how the tuning improves the current regular process.

Hardware

The hardware used in this research was composed of eight drones DJI Phantom 3 Standard, eight MacBook Pro 2015 with an Intel Core i5 processor, 16 GB RAM and 2 GB video GPU each. Each drone took videos in one of these cities: (1) Caracas, Venezuela; (2) Maracaibo, Venezuela; (3) Cancún, Mexico; (4) Santo Domingo, Dominican Republic; (5) Bogota, Colombia; (6) Quito, Ecuador; (7) Sao Paulo, Brazil; and (8) Buenos Aires, Argentina. The transmission used an independent program developed by the author. Drone videos were sent in divided frames.

Each MacBook was aligned with a drone to send transformed frames to the public cloud, specifically an AWS account. The author used an external storage device with each MacBook. Eight devices detected DJI radio frequency with eight Arduino boards.

Concerning the public cloud, the author used three to eight t2.xlarge Elastic Compute Cloud (EC2) instances with four virtual central processing units (vCPUs) and 16 gigabyte (GB) of random access memory (RAM) using Amazon Machine Images (AMIs). The complete list of EC2 instances in AWS can be found online (www.aws.amazon.com).

Software

The software/services developed and used by the author includes:

- **DRON2PREM:** This software was developed in C language by the author of the article to read every frame from the drone using a live streaming feature. It is automatically saved in an external

device connected to one MacBook Pro. This software decodes the drone's video frames with a wireless sensor at the same radio frequency.

- **PREM2CLOUD:** This software was developed in C language by the author of the article to run at the same time as DRON2PREM. This program compresses the final videos with 7z. This process may be slower than other compress methods due to the final file size. After compressing it in multiple parts, it uses AWS Command Line Interface (CLI) to simultaneously send files with several threads. The files are allocated in a Simple Storage Service (S3) bucket.
- **CLOUD2VALUE:** This software was developed by the author of the article to analyze data on the cloud. The program, which was written in Python language, transform every drone video frames file allocated in the S3 bucket to XML files and processes every XML to search for patterns. This software can use some machine learning processes and detects desired patterns. For example, in this article, it is used to review the areas searching for red points.

The software/services developed by external companies include:

- **Amazon Web Services:** These services were used to develop experiments. The author used several AWS services, including Simple Storage Services (S3), Elastic Load Balancer (ELB), and Autoscaling.
- **Tsunami UDP:** This service is a “fast user-space file transfer protocol that uses TCP control and UDP data for transfer over very high speed long distance networks (³ 1 Gbps and even 10 GE), designed to provide more throughput than possible with TCP over the same networks” (Wagner, 2009, para. 1).

IMPLEMENTATION

The project's implementation includes complex steps. In the first part, the author creates a program with Python, C, and Bash programming languages to obtain data from DJI drone transmission. Radio frequency sensor device technology was required to decode data frames sent by drones. These devices were created locally using Arduino boards. They were connected to the machines via a universal serial bus (USB) port. An instance of the DRON2PREM program runs in the local machine. It saves the drone's data from the Arduino boards. These data frames are saved as a compressed file inside an 8 TB external device.

At the same time, an instance of PREM2CLOUD is running. When DRON2PREM has written the final message in the current log file of the current thread, the PREM2CLOUD instance divides the file into eight parts. This creates a multi-parallel connection with the AWS S3 bucket to send parts via channel created with different network interface cards.

There is a listener waiting for new connections in the CLOUD2VALUE program in an EC2 instance in AWS platform. When a new connection arrives, the CLOUD2VALUE software waits to finish the message in the transfer log file of the PREM2CLOUD. After the PREM2CLOUD finishes the copy from an on-premise machine to the S3 bucket, the CLOUD2VALUE link the eight pieces of the compressed video. The CLOUD2VALUE will uncompress the file and place it in a local (for the EC2 instance) elastic block storage (EBS) device for the hard process.

The CLOUD2VALUE program reads the video frames. For every frame, it produces an XML file with color information on every pixel of the frame. The author used the red-green-blue (RGB) system

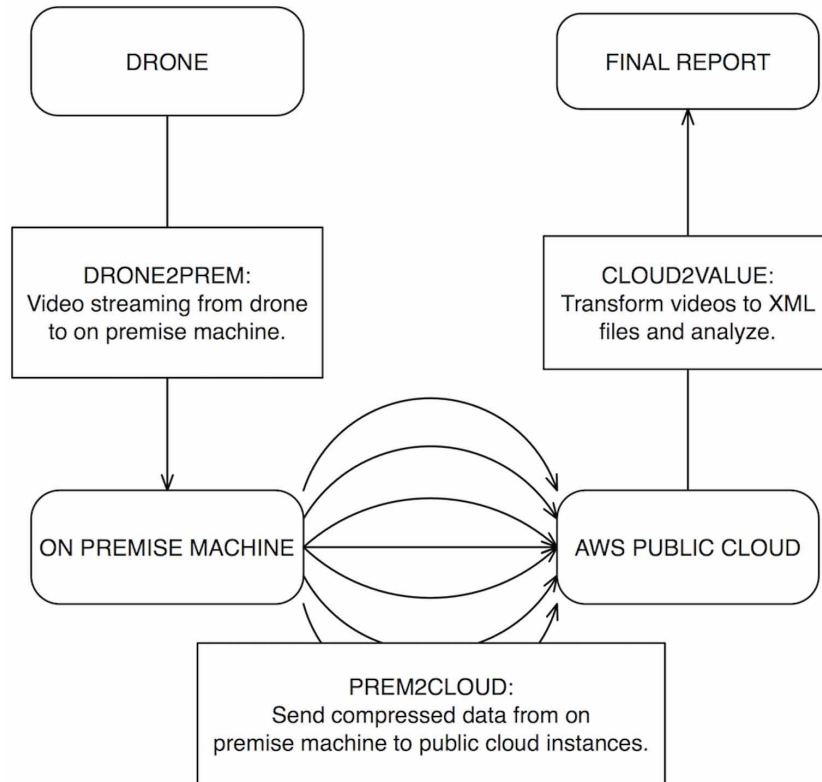
to process pixel colors. After every frame is processed, the author's neural network (NN) Drone Image Processor (DrImPro) program will search the requested input.

The experiment includes auto scaling, which is a new feature of the public cloud. This feature balances the process between multiple nodes (EC2 instances). If necessary, it creates new EC2 instances to help in the process. For example, if the current project has three EC2 instances with all instances using 100% CPU, the auto scaling feature can create new EC2 instances to balance the processing load.

Another test was created with experimental drones (different from DJI). This test evaluates the transmission between the drone and the on-premise machine with the DRON2PREM (not restricted to DJI brand drones). For these tests, the author used Raspberry Pi boards connected to visual sensors to construct the drones.

At the end of the experiment, the program reveals the summary of every node in the machine learning (ML) network and the process. Figure 1 illustrates the global process.

Figure 1. Global process of the tuning and analyzing portions



RESULTS

The experiments selected for this article are:

- Find the areas with more red points (basic experiment).
- Find the areas with more blue points (simulating water).
- Find the areas with more cars (simulating traffic).

The researcher ran the experiment in different combinations to demonstrate the hypothesis of the improvement of the process with the new method. The main combinations are:

1. Eight DJI drones, eight MacBook Pro laptops, on-premise NN processing
2. Eight DJI drones, eight MacBook Pro laptops, one copy thread per laptop, three to eight EC2 instances as nodes in the NN
3. Eight DJI drones, eight MacBook Pro laptops, eight copy threads per laptop, one EC2 instance as nodes in the NN
4. Eight experimental drones, eight MacBook Pro laptops, eight copy threads per laptop, three to eight EC2 instances as nodes in the NN
5. Eight DJI drones, eight MacBook Pro laptops, eight copy threads per laptop, three to eight EC2 instances as nodes in the NN

Every test uses an average of 1.5 TB of raw data in total. When transformed to XML files, the size was reduced to an average of 18 GB. Using those experiments, the author can prove the method's improvement in translating the processing to the cloud and using auto scaling to benefit research.

Figure 2 illustrates the results. The first experiment used eight DJI drones; it executed three programs (i.e., DRON2PREM, PREM2CLOUD, and CLOUD2VALUE) in on-premise machines. They saved compress time, decompress time, and time spent sending the video to the public cloud. However, the difference in time was not significant because it used a local network with less bandwidth than the public cloud network. The experiment required a huge amount of costly resources. The experiment's result values were 15 hours, 27 minutes, and 13 seconds.

In the second experiment, the author aimed to show slow processing when using only one thread, which is the common technique. However, the author did not employ auto scaling. Without auto scaling, the time to complete the task was very high (greater than five days). Time was invested in the network when transmitting compressed files. The total time for the second experiment was 39 hours, 15 minutes, and 35 seconds. This employed a 1 GB/s network link.

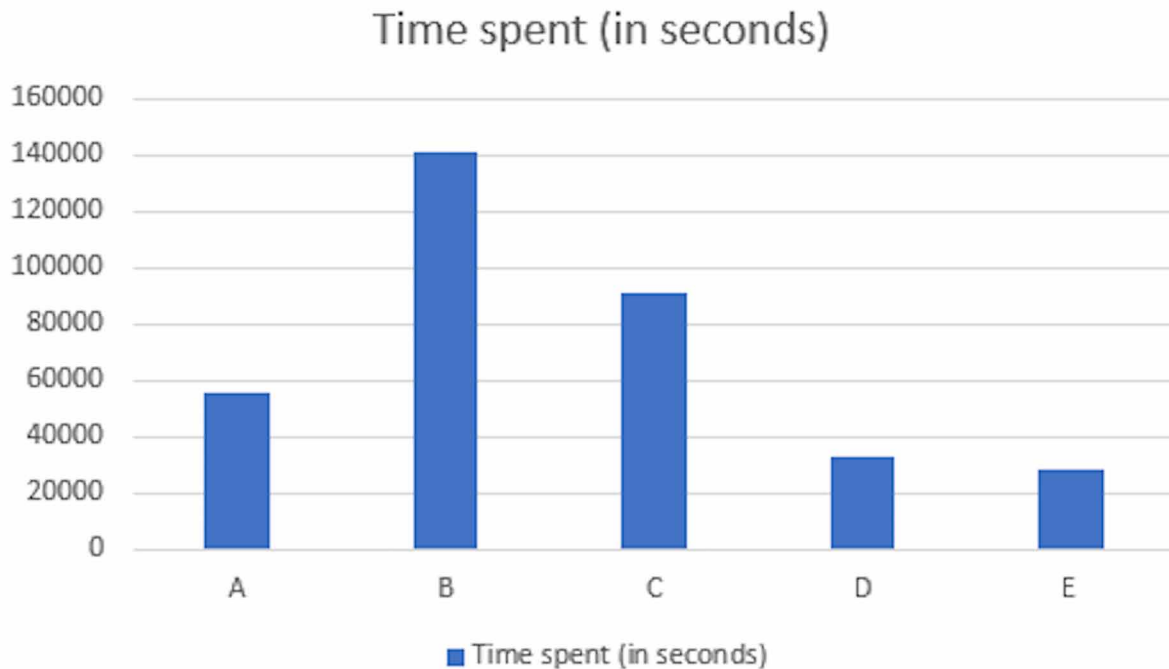
On the other hand, the third experiment demonstrated the value of auto scaling. In this example, the author used only one EC2 instance in the public cloud to process all data. The time was high (25 hours, 19 minutes, and zero seconds).

In the fourth experiment, the researcher aimed to experiment with custom programs and custom drones created by electronical and computer engineers. They used Raspberry Pi boards as the core body of the drones. Several sensors (video, Bluetooth, and GPS) connected to perform the experiment. The results were impressive as the process was completed in nine hours, 11 minutes, and five seconds with the use of eight threads of network transmission and auto scaling.

In the final experiment, the author used their best resources, including eight DJI drones, eight laptops as local receivers, eight threads of network copy, and three to eight EC2 instances using auto scaling on the AWS public cloud. With these parameters, the complete process was executed in eight hours, one minute, and five seconds.

With the results of this study, companies and regular researchers can have a new alternative to improve their transference process of drone videos (or another types of videos) to be processed on the public cloud. Also, they have a better method of processing data using the Autoscaling feature of Amazon Web Services. With this advanced method, the process of analyzing data from drone videos on the public cloud reach a 78% of improvement. This is considerable time to spent in other tasks more related with the interpretation of the results or the output.

Figure 2. Comparison of time invested in complete processing using different methods



SOLUTIONS AND RECOMMENDATIONS

During the experimental evaluation, the researcher must confirm that other drones do not exist in the UAV area due to the risk of failed transmission. The reception of drones' transmission packets can capture packets from other drones in the air. This will result in a failed experiment. A solution for this situation is send the drone packets in a different radio frequency and another channel instead of the default of the brand, only if this is a permitted behavior of the hardware involved. For this experiment, the author checked the area before, during, and after the experiment. Other drones were not present within 500 meters of the round ratio. Therefore, the experiment was highly reliable.

Another recommendation for new experiments is to use a virtual private network (VPN) for transmission from on-premise machines to the public cloud. VPNs allow for better performance and security in the transport of drone data. This article's experiments did not use security cryptographic algorithms or keys. The author used the default device environments.

FUTURE RESEARCH

Using this article, researchers who are interested in drones and public cloud development can obtain valuable resources to expand tests. For example, this study did not cover the scope of different drone brands (i.e., Blackhawk, Altar) or public cloud platforms (i.e., Google Cloud Platform, Microsoft Azure). An extension of this study can compare transmissions using different platforms as variables. Future studies can evaluate which platform is best for neural network or machine learning with drone image processing.

The author is working on an extension of this study. It will include the development of a mechanism to detect a specific face among hundreds of individuals. As the drone flies and sends image frames to the station and the public cloud for processing, the cloud program will process images. Along with the machine learning and big data processing, the program should provide specific coordinates to locate the person in the picture.

To review the software in Python or C used, they can be found in the webpage of the project, listed in the references section. It uses a GNU General Public License (GPL) for the source code of every component.

CONCLUSION

In this chapter, the author created a complex way to optimize the processing of long drone videos on the public cloud using several GB or TB per video. The researcher's system collects a drone-captured video, sends the data to a local on-premise machine, tunes transmission of the video to the public cloud, transforms the videos to XML files, and creates a custom program to search and find points of interest in every frame to indicate a specific point. The author tuned the normal process an average of 78% of the time.

The results in this study have several practical implications, for instance, they can be very useful in academic researches about environment. The process of analyzing biological, weather or environment data from on premise machines to the cloud is expensive and slow. Another example could be the process of rendering and editing movies or television series on the cloud. When the team record the videos and the platform or application is allocated on the cloud, they have to send all the scenes to the platform. After that, they can edit the videos. With this new method, the process would improve in cost and time.

REFERENCES

- Bitam, S. (2012). *Bees Life Algorithm for Job Scheduling in Cloud Computing*. Paper presented at the 15th International Conference on Computer and Information Technology (ICCIT), Hammamet, Tunisia.
- Drone. (2018). *OxfordDictionaries.com*. Retrieved from <https://en.oxforddictionaries.com/definition/drone>
- Gharibi, M., Boutaba, R., & Waslander, S. (2016). Internet of drones. *IEEE Access: Practical Innovations, Open Solutions*, 4, 1148–1162. doi:10.1109/ACCESS.2016.2537208
- Google Cloud Platform. (2017). *GCP Storage Service*. Retrieved from <https://cloud.google.com/storage/>
- Koubaa, A., Qureshi, B., Sriti, M., Javed, Y., & Tovar, E. (2017). *A service-oriented cloud-based management system for the internet-of-drones*. Paper presented at the CISTER Research Centre, Porto, Portugal. 10.1109/ICARSC.2017.7964096
- Liu, B., Chen, Y., Blasch, E., Pham, K., Shen, D., & Chen, G. (2014). A Holistic Cloud-Enabled Robotics System for Real-Time Video Tracking Application. In *Future Information Technology*. Springer-Verlag. doi:10.1007/978-3-642-40861-8_64
- Lo, J. (2018). *Improving analysis of drone videos on the cloud*. Retrieved from <http://www ldc.usb.ve/~jose/dronecloud/index.html>

Lo, J., Lopez, H., Goncalves, M., & Perera, G. (2015). Study of Skyline Query Evaluation on Corona. In *Encyclopedia of information science and technology* (3rd ed., Vol. 3, p. 182). IGI Global. doi:10.4018/978-1-4666-5888-2.ch182

Mahmoud, S., Mohamed, N., & Al-Jaroodi, J. (2015). Integrating UAVs into the cloud using the concept of the Web of things. Hindawi. *Journal of Robotics*, 631420.

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology* [Special Publication 800-145]. National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-145

Photogrammetry. (2018). *OxfordDictionaries.com*. Retrieved from <https://en.oxforddictionaries.com/definition/photogrammetry>

Saleem, Y., Salim, F., & Rehmani, M. (2014). Resource Management in Mobile Sink Based Wireless Sensor Networks through Cloud Computing. In *Resource Management in Mobile Computing Environments, Modeling and Optimization in Science and Technologies 3*. Springer International Publishing. doi:10.1007/978-3-319-06704-9_20

Sattar, N., Adnan, M., & Kali, M. (2017). *Secured Aerial Photography Using Homomorphic Encryption*. Paper presented at the 2017 International Conference on Networking, System and Security (NSysS), Dhaka, Bangladesh. 10.1109/NSysS.2017.7885810

Wagner, J. (2009). *Tsunami UDP protocol*. Retrieved from <http://tsunami-udp.sourceforge.net/>

Welsh, R. (2016). Utilizing Massive Compute Resources in the Public Cloud for Complex Image Processing. *SMPTE Motion Imaging Journal*, 125(8), 38–48. doi:10.5594/JMI.2016.2602122

Yinong, Z. D., & Garcia, M. (2010). *Robot as a service in cloud computing*. Paper presented at the 5th IEEE International Symposium on Service Oriented System Engineering (SOSE), Nanjing, China.

ADDITIONAL READING

Erl, T. (2005). *Service-oriented architecture: Concepts, technology and design*. Prentice Hall.

Erl, T., Mahmood, Z., & Puttini, R. (2014). *Cloud computing: Concepts, technology & architecture*. Prentice Hall.

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645–1660. doi:10.1016/j.future.2013.01.010

Mahmoud, S., & Mohamed, N. (2014). *Collaborative UAVs cloud*. Paper presented at the International Conference of Unmanned Aircraft Systems (ICUAS). 10.1109/ICUAS.2014.6842275

NASA. (2015). *NASA UTM 2015: The next era of aviation*. Retrieved from <https://utm.arc.nasa.gov/utm2015.shtml>

Nolan, M. (2010). *Fundamentals of air traffic control* (5th ed.). Delmar.

Stallings, W. (2005). *Wireless communications & networks* (2nd ed.). Prentice-Hall.

Zeng, D., Guo, S., & Cheng, Z. (2011). The Web of things: A survey. *Journal of Communication*, 6(6), 424–438.

KEY TERMS AND DEFINITIONS

Autoscaling: Automatic load balance service and instance creator software designed by Amazon Web Services.

Compress: Reducing the size of a computer file or object using compression algorithms.

Drone: A remote robot device that can fly and execute specific programmed activities.

Machine Learning: The process in which a computer learns a little bit in every iteration of an algorithm.

Neural Network: A segregated computer system program that uses example points and processes a set of data based on these points to give specific results.

On Premise: A set of physical resources on the building or site where the person who is executing the program is located.

Parallel Threads: Several lines of processing used in a big data load.

Public Cloud: On-demand computing services provided by enterprises with huge amounts of resources.

Tuning: Improving a process or execution of a task.

Improving the Integration of Distributed Applications

2

José Carlos Martins Delgado

 <https://orcid.org/0000-0002-2536-4906>

Instituto Superior Técnico, Universidade de Lisboa, Portugal

INTRODUCTION

Complex software systems typically involve many interacting applications, with many decisions to take and many tradeoffs to evaluate, not only in each application but also in the ways in which the various applications interact. Ideally, applications should be completely decoupled, with no constraints on one another and completely independent lifecycles. This would allow separate development of each application and elimination of design and programming inefficiencies due to interaction between the specifications of applications, which usually cause iterations in requirements for other applications and consequent changes.

However, applications do need to interact and to cooperate, to collectively fulfill the goals of the system. Therefore, a fundamental tenet in software design is to reduce application coupling as much as possible (Bidve, & Sarasu, 2016) without hindering the interaction capabilities necessary to support the required application interoperability. Application decoupling has also the advantage of improving:

- § **Changeability:** A change in one application is less likely to have a significant impact in other applications.
- § **Adaptability:** Fewer constraints require less effort to adapt to changes in other applications.
- § **Reusability:** An application with fewer requirements and constraints has an increased applicability range.
- § **Reliability:** A smaller set of requirements simplifies the task of finding an alternative application in case of failure.

In general, the fundamental problem of application design, in terms of interaction, is how to provide (at most) the minimum coupling possible while ensuring (at least) the minimum interoperability requirements. Therefore, the main goal is to ensure that each interacting application knows just enough about the others to be able to interoperate with them but no more than that, to avoid unnecessary dependencies and constraints. This is an instance of the principle of least knowledge (Hendricksen, 2014).

The two most used application integration approaches, Service-Oriented Architecture (SOA) (Erl, Merson, & Stoffers, 2017) and Representational State Transfer (REST) (Fielding, Taylor, Erenkrantz, Gorlick, Whitehead, Khare, & Oreizy, 2017) hardly comply with the principle of least knowledge. They achieve interoperability but do not solve the coupling problem, since they require, in practice, that the schemas used by the interacting applications are the same.

SOA is good at modeling distributed systems based on the service paradigm (an extension the object-oriented paradigm for distributed applications) but involves rather complex and static service specifications and entails sharing schemas between client and server, which is a heavy form of coupling. Changing the

DOI: 10.4018/978-1-7998-3479-3.ch017

interaction between Web Services is not a trivial task. REST is much simpler, justifying its increasing popularity, but is rather low level and not the best match for general-purpose, behavior-oriented distributed applications. It also entails a high level of coupling, since it requires that both interacting applications share the same media type specification.

This article revisits the application integration problem with an open mind, without *a priori* restrictions from specific technologies, such as Web Services for SOA and RESTful APIs for REST. The only assumption is that there are applications that need to interact, by using messages. The idea is to base the interoperability mechanism on the concepts of *compliance* (Czepa, Tran, Zdun, Kim, Weiss, & Ruhsam, 2017) and *conformance* (Carmona, van Dongen, Solti, & Weidlich, 2018), which allow partial interoperability, rather than on sharing data schemas. This makes all the difference. Unlike Web Services, there are no declared schemas that a client is forced to use in their entirety. Unlike RESTful applications, the client and server do not have to agree on a specific data type. Each resource (dataset, message or distributed application) has its own schema, stemming from its service (interface) definition. Checking for interoperability between two resources (for example, between a message and the parameter required by an operation) is done structurally, component by component, recursively until primitive resources are found.

The rest of the article is structured as follows. It starts by establishing a simple interoperability framework, with the various abstraction levels at which interoperability must be established. After defining the fundamental problem of application integration, the two most common solutions, SOA and REST, are analyzed and compared, in particular in what coupling is concerned. Compliance and conformance are then presented and proposed as a solution to reduce coupling to the bare minimum required by the interacting applications.

BACKGROUND

Modern applications, designed and managed in a distributed way, need to interact and collaborate with an increasing scale, since systems are becoming more complex and diversified. Digital-based technologies, such as cloud computing (Varghese, & Buyya, 2018), mobile computing (Page, & Thorsteinsson, 2018), and the Internet of Things (Paul, & Saraswathi, 2017) became ubiquitous and disruptive. In the recent 4.0 trend (Dornberger, 2018), collaboration means generating and exchanging more and more data, either at business, personal, or sensor levels. The fourth industrial revolution, commonly known as Industry 4.0 (Liao, Deschamps, Loures, & Ramos, 2017), in which agile reconfigurability of the production supply chain is a fundamental objective, is just an example.

All this raises the application integration problem to a completely new level, in which conventional integration technologies expose their limitations and require new solutions.

Integration (Panetto & Whitman, 2016) can be broadly defined as the act of instantiating a given method to design or adapt two or more systems, so that they cooperate and accomplish one or more common goals. To interact, applications must be interoperable, i.e., able to meaningfully operate together.

The ISO/IEC/IEEE 24765 standard (ISO, 2010) defines *interoperability* (Agostinho, Ducq, Zacharewicz, Sarraipa, Lampathaki, Poler, & Jardim-Goncalves, 2016) as “the ability of two or more systems or components to exchange information and to use the information that has been exchanged”. Therefore, merely exchanging information is not enough. Interacting applications must also be able to understand it and to react according to each other’s expectations.

A related problem is *coupling* (Bidve, & Sarasu, 2016), which provides an indication of how much applications are intertwined and depend on each other. Some degree of coupling is unavoidable, since

some form of mutual knowledge is necessary to make interoperability possible. Interoperability and a coupling as low as possible need to be combined to achieve an effective and efficient integration of distributed applications.

The two most used integration approaches are the Software-Oriented Architecture (SOA) (Erl, Merson, & Stoffers, 2017) and Representational State Transfer (REST) (Fielding, Taylor, Erenkrantz, Gorlick, Whitehead, Khare, & Oreizy, 2017), with the corresponding technological solutions for distributed interoperability, Web Services (Zimmermann, Tomlinson, & Peuser, 2012) and RESTful APIs (Pautasso, 2014), respectively.

Unfortunately, the discussion involving SOA and REST (Bora & Bezboruah, 2015) usually entails more technological issues than conceptual or modeling arguments. There are also several proposals to integrate SOA and RESTful services (Dahiya, & Parmar, 2014; Sungkur, & Daiboo, 2016).

Web Services and RESTful APIs are based on data description languages such as XML (Fawcett, Ayers, & Quin, 2012) and JSON (Bassett, 2015). Although these languages have achieved the basic objective of interconnecting independent and heterogeneous systems, supporting distributed application interoperability, they are not effective solutions from the coupling viewpoint, since they require that the data schemas of the messages exchanged by the interacting applications are the same.

Several metrics have been proposed to assess the maintainability of service-based distributed applications, based essentially on structural features, namely for service coupling, cohesion and complexity (Babu, & Darsi, 2013). There are also approaches trying to combine structural coupling with other levels of coupling, such as semantics (Alenezi, & Magel, 2014).

Compliance (Czepa, Tran, Zdun, Kim, Weiss, & Ruhsam, 2017) is a concept that can be used as the foundation mechanism to ensure partial interoperability and thus minimize coupling. It has been studied in specific contexts, such as choreography (Capel & Mendoza, 2014), modeling (Brandt & Hermann, 2013), programming (Preidel & Borrmann, 2016), and standards (Graydon, Habli, Hawkins, Kelly, & Knight, 2012). Conformance (Carmona, van Dongen, Solti, & Weidlich, 2018) is another concept underlying partial interoperability, enabling an application to replace another if it conforms to it (supports all of its features).

A Simple Interoperability Framework

Most interoperability technologies, such as Web Services (Zimmermann, Tomlinson, & Peuser, 2012) and RESTful APIs (Pautasso, 2014), usually consider only the syntactic format of messages, or at most the semantics of its terms (Wang, Gibbins, Payne, Patelli, & Wang, 2015). However, the ability to meaningfully interact involves higher levels, naturally including behavior. Table 1 establishes a linearized systematization of interoperability at several levels of abstraction. The higher levels are particularly relevant when complex, enterprise-class applications are involved.

All these abstraction levels are always present in all application interactions, even the simplest ones. There is always a motivation and purpose in sending a message, an effect stemming from the reaction to its reception, a meaning expressed by the message, and a format used to send it over a network under some protocol. However, what happens in practice is that some of these levels are catered for *tacitly* (implicitly assumed) or *empirically* (explicitly hidden underneath some existing specification).

Also relevant is non-functional interoperability. It is not just a question of invoking the right operation with the right parameters. Adequate service levels, context awareness, security and other non-functional issues must be considered when applications interact, otherwise interoperability will be less effective or not possible at all. The abstraction levels of Table 1 must be considered for each of these concerns.

Table 1. Abstraction levels and layers of interoperability

Abstraction level	Layers	Description
Symbiotic (purpose and intent)	§ Coordination § Alignment § Collaboration § Cooperation	§ Motivations to interact, with varying levels of mutual knowledge of governance, strategy and goals.
Pragmatic (reaction and effects)	§ Contract § Workflow § Interface	§ Management of the effects of the interaction at the levels of choreography, process and service.
Semantic (meaning of content)	§ Inference § Knowledge § Ontology	§ Interpretation of a message in context, at the levels of rule, known application components and relations, and definition of concepts.
Syntactic (notation of representation)	§ Structure § Predefined type § Serialization	§ Representation of application components, in terms of composition, primitive components and serialization format of messages.
Connective (transfer protocol)	§ Messaging § Routing § Communication § Physics	§ Lower level formats and network protocols involved in transferring a message from the context of the sender to that of the receiver.

The Fundamental Problem of Application Integration

Applications that interact in a distributed environment require some network to send each other messages. In each message transaction (sending a request and processing a response), both sender and receiver need to understand and react appropriately upon receiving request and response messages, at all abstraction levels of Table 1.

A given application, in the role of server, publishes the set of request messages to which it is able to respond, which defines the interface of the functionality offered by that application (its Application Programming Interface – API). In the role of client, another application may send one of the acceptable request messages to the server and invoke the corresponding functionality. Applications can also express their APIs in terms of exposed features, such as operations. Each operation can accept its own set of request messages.

The server needs to be able to understand the client's request and to react and respond according to what the client expects. Otherwise, the interaction will not produce the intended effects.

Coupling expresses the mutual dependencies between applications, stemming from the need to understand messages. The goal is to reduce it as much as possible, to avoid unnecessary constraints to the evolution and variability of applications. However, what can be achieved is a balance between two contradictory goals:

- § On the one hand, two completely uncoupled applications can evolve freely and independently, which favors adaptability, changeability and even reliability (if one fails, there is no impact on the other).
- § On the other hand, applications need to interact to cooperate towards common or complementary goals, which means that some degree of previously agreed mutual knowledge must exist. The more they share with the other, the easier interoperability gets, but the greater coupling becomes.

The fundamental problem of application integration is how to provide *at most* the minimum coupling possible, while ensuring *at least* the minimum interoperability requirements. The main goal is to ensure

that each interacting application knows just enough about the others to be able to interoperate with them but no more than that. Unnecessary dependencies and constraints should be avoided.

Existing data interoperability technologies, such as XML and JSON, assume that interacting applications share the same message schema. This is reminiscent of the first days of the Web and HTML documents, in which the client read the document produced by the server and both needed to use the same data specification. Today, data have been separated from their specification (schema), but both client and server still work on the same information. Decoupling has not improved.

The problem is that a server may need to serve several potentially different clients and a client may need to send requests to several potentially different servers. Sharing a schema couples both client and server for the whole set of messages that the schema can describe, even if the client only uses a subset of the admissible requests and the server only responds with a subset of the admissible responses. Client and server are therefore more coupled than needed and changes in one application may very likely imply changing the other as well, even if a change does not affect the actually exchanged messages.

Current Solutions to the Application Integration Problem

SOA (Erl, Merson, & Stoffers, 2017) and REST (Fielding, Taylor, Erenkrantz, Gorlick, Whitehead, Khare, & Oreizy, 2017) are the most popular architectural styles today in the integration of distributed applications.

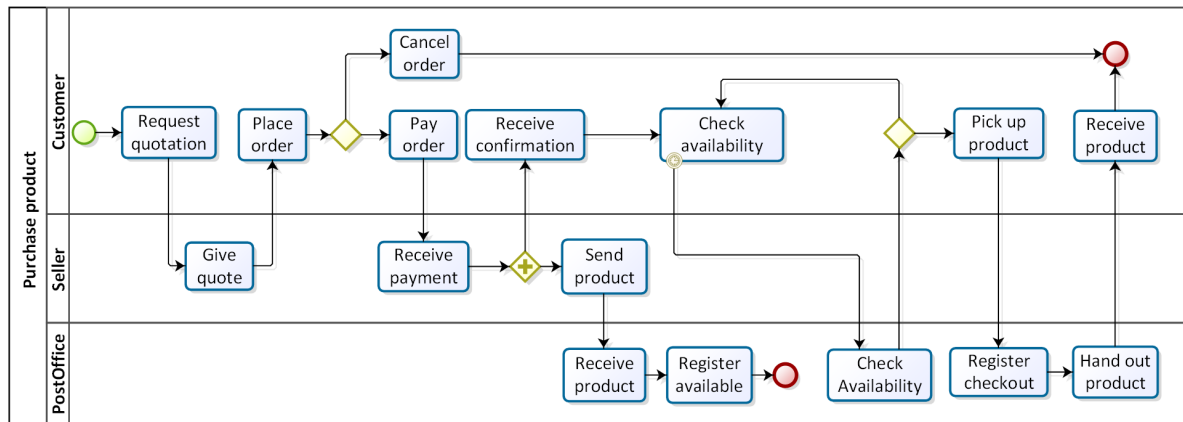
SOA appeared in the context of enterprise integration, with the main goal of achieving interoperability between existing enterprise applications with the then emerging XML-based technologies, more universal than those used in previous attempts, such as CORBA (Common Object Request Broker Architecture) (Henning, 2008) and RPC (Kukreja & Garg, 2014). It was a natural evolution of the object-oriented style, now in a distributed environment and with large-grained resources to integrate. Therefore, the emphasis was put on modeling applications as black boxes, exposing their external functionality with an interface composed of a set of operations, application dependent. Internal structure was avoided, as a measure to reduce coupling by applying the information hiding principle.

REST appeared in the context of Web applications, as a systematization of the underlying model of the Web and with scalability as the most relevant goal, since interoperability had already been solved by HTML (and, later, XML and JSON) and HTTP. REST is based on several architectural constraints, the most relevant of which are stateless interactions (only the client maintains the interaction state) and uniform interface (all resources have a service with the same set of operations, with variability on structure and with links to resources, rather than on service interface). These constraints have decoupling between a server and its clients as their main motivation, to support scalability. When potentially thousands of clients connect to a server, server applications need to be able to scale and to evolve as needed and as independently as possible from the clients.

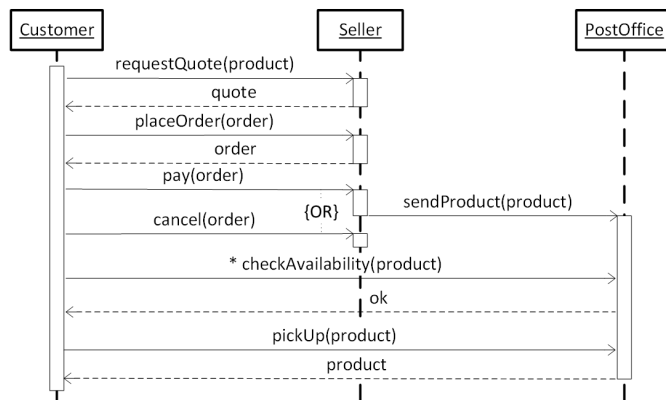
Coupling was therefore a main concern for both architectural styles, although this has not prevented a lively debate between proponents of SOA and REST. In particular, REST has been heralded as much simpler and providing a better decoupling between interacting applications (Pautasso, Zimmermann, & Leymann, 2008). A simple example, however, indicates that these two architectural styles are not that different, in essence.

Figure 1a describes a typical online product purchase scenario, using the process paradigm and Business Process Modeling Notation (BPMN) (Chinosi & Trombetta, 2012) to illustrate the interactions between a customer, a seller, and a logistics company to deliver the product.

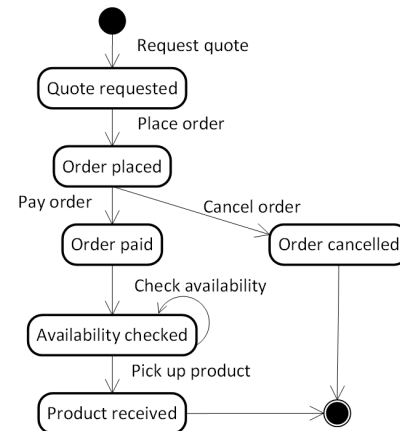
Figure 1. An example of business-level distributed interactions (a), with a model of the customer process in (b) SOA and (c) REST styles



(a)



(b)



(c)

Each of the three roles in this process choreography needs to be conceived, designed and implemented as much independently as possible, since each role can be performed by applications with different goals and characteristics, which means that coupling should be minimized.

This example can be used to compare the approaches based on SOA and REST. Figure 1 details just the Customer process, for simplicity, using UML. Figure 1b represents the SOA approach, with a resource structure mimicking the three actors and the corresponding sequence diagram, reflecting their interactions. The REST approach goes directly to modeling interaction states, which is reflected in Figure 1c by a state diagram.

Apart from details, it is noticeable that the flow of activities of the customer in SOA is dual of the flow of states in REST. Figure 1b (SOA) uses an activity-based flow graph, with activities performed at the nodes and state changed at the edges, whereas Figure 1c (REST) places state at the nodes and implements activities at the edges.

SOA is guided by behavior and REST by (representation of) state. In UML terminology, SOA uses a static class diagram and a sequence diagram (Figure 1b) as a first approach, to specify which resource types are used, to establish the set of operations provided by each, and their interactions, whereas REST

starts with a state diagram (Figure 1c), without relevant concern about distinguishing which state belongs to which resource. In the end, they perform the same activities and go through the same states, although with different perspectives that stem from the approaches used to solve the same problem.

It is also clear that both SOA and REST tackle the fundamental problem of application integration, to achieve interoperability while trying to minimize coupling and to comply with the information hiding principle by hiding the implementation of individual resources:

- § SOA's approach is to hide resource structure (considering it part of the implementation), whereas REST's approach is to hide resource behavior (its service) by converting it into resource structure (each operation is modeled as a resource).
- § SOA models resources as close as possible to real world entities, to reduce the semantic gap (Sikos, 2017). Since these are different from each other, offering a specific set of functionalities, the result is a set of resources with different interfaces (different set of operations). The client must know the specific interface of the server. This entails coupling and may only be acceptable if the number of clients for a given server is usually reasonably small and the system evolves slowly (to limit the maintenance effort).
- § REST tries to avoid the interface coupling by decomposing complex resources into smaller ones (with links to other resources) until they are so primitive that they can all be treated in the same way, with a common interface. The interface diversity is transferred to the richness of the structure of the links between resources. This uniform interface for all resources, with a common set of operations, is the most distinguishing feature of REST. This corresponds to separating the mechanism of traversing a graph (the links between resources) from the treatment of each node (resource). The goal is that a universal link follow-up mechanism, coupled with a universal resource interface, leads to decoupling between server and clients, allowing servers to change what they send to the clients because these will adapt automatically, by just navigating the structure and following the links to progress in the interaction process.

Unfortunately, both SOA and REST have fallacies and limitations:

- § SOA assumes that all resources (applications) are at the same level and that exposing their services is just providing some interface and the respective endpoint. The fallacy is to assert that resource structure is unimportant. This may be true when integrating very large-grained resources, but not in many applications that include lists of resources that need to expose their services individually. The limitation is in changeability. Changing the interface of the server will most likely require changes in the clients as well.
- § REST imposes a uniform interface, which allows treating all the resources in the same way. The fallacy is to consider only a syntactic interface, or even semantic (Verborgh, Harth, Maleshkova, Stadtmüller, Steiner, Taheriyani, & Van de Walle, 2014), but forgetting behavior (higher levels in Table 1). The reaction of resources to requests also needs to be considered, which means that following a link cannot be done blindly. The client needs to know which kind of reaction the server is going to have. In the same way, traversing a structured resource implies knowing the type of that resource, otherwise which link should be followed next may not be known. This means that all types of resources to be used must be known in advance, either standardized or previously agreed upon (with custom media types). This is a relevant limitation of REST. What happens in practice is that developers are more than happy to agree on simple resource structures and on an aligned

API to match. REST's advocates contend that an API should be represented simply by its initial link (an URI – Universal Resource Identifier –, for example) and all the functionality should be dynamically discovered by the client by exploring links. However, this can only be done with a fixed set of resource types (predefined or previously agreed upon). Resource structure can vary, a good advantage of REST, but only within the boundaries of resource types known in advance. This is nothing more than sharing schemas, which means coupling.

SOA is a better option for large-grained, slowly-evolving complex resources, whereas REST is a better option for small-grained, structured resources that are relatively simple. In search for simplicity and maintainability, many service providers (including cloud computing providers) have stopped using SOA APIs in favor of REST, by modeling operations as resources, using naturally occurring lists of resources as structured resources with links, and dynamically building URIs instead of resorting to operation parameters. Decoupling in REST, however, is not as good as its structure dynamicity seems to indicate. All the resource types (media types, in REST terminology) must be known in advance and the client may break if a new media type unknown to the client is used by the server. In addition, even when the type is known, the client cannot invent what to do and which link to choose when it receives a resource representation of a given type. Collaboration requires that both server and client implement behavior and deal with its effects, which means that they cannot be agnostic of each other, but rather need to be designed to work together.

Minimizing Coupling With Compliance and Conformance

With current integration technologies, interacting applications must share the schemas of the messages that they exchange. This leads to a higher level of coupling than actually needed, since applications are constrained by the whole set of features of those schemas, even if not all are used.

This article proposes an alternative approach, in which the schema of a message, as generated by the sender, needs only partially match the schema of the messages expected by the receiver. Only the actually used features need to have a match. Features present in the message but not required by the receiver are ignored. Optional receiver features not present in the message are assigned default values prior to processing the message at the receiver. In addition, structured features are checked structurally and recursively, component by component.

This means that a client can use (send request messages to) various servers, as long as these messages match the relevant parts of those servers' request schemas, even if they were not designed to work together. In addition, a server can be replaced by another one, with a different schema, as long as it can deal with all the requests that the replaced one could. The replacement can occur due to evolution of the server to a new version or by resorting to another, similar server, probably for load balancing.

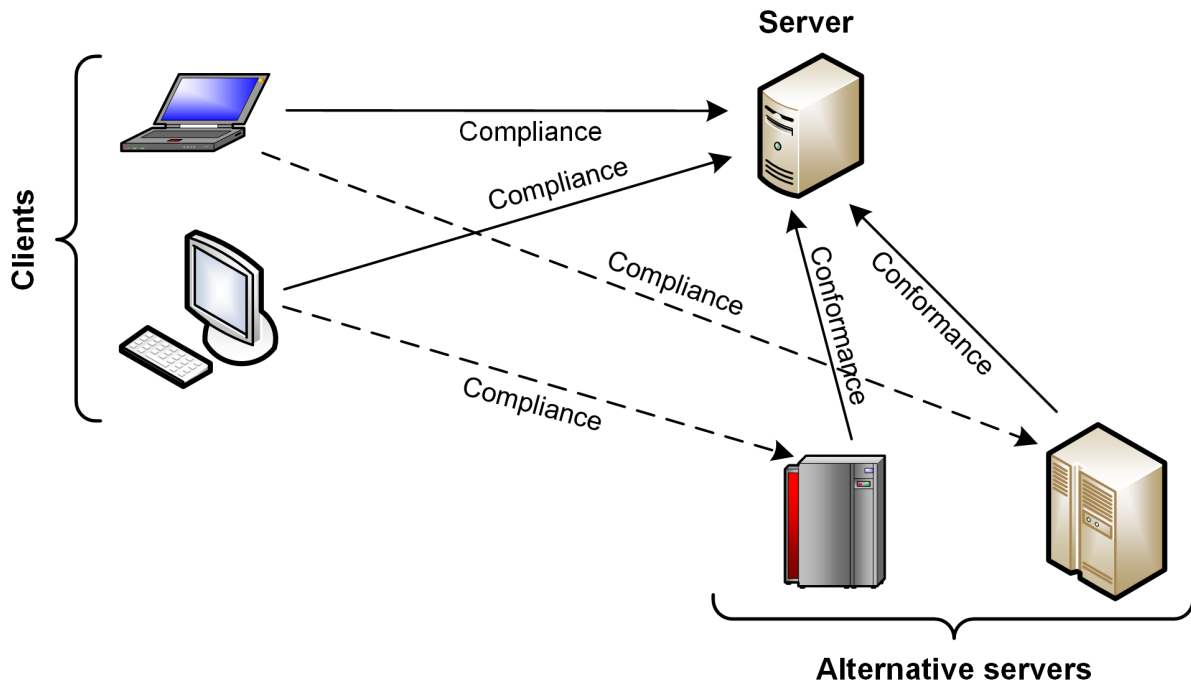
These *use* and *replace* relationships lead to two important schema relations, illustrated by Figure 2:

- § **Client-server: Compliance** (Czepa, Tran, Zdun, Kim, Weiss, & Ruhsam, 2017). The client must satisfy (*comply with*) the requirements established by the server to accept requests sent to it, without which these cannot be validated and executed. It is important to note that any client that complies with a given server can use it, independently of having been designed for interaction with it or not. The client and server need not share the same schema. The client's schema needs only to be compliant with the server's schema in the features that it actually uses. Since distributed applications have independent lifecycles, they cannot freely share names, and schema compliance must

be checked structurally, feature by feature, between messages sent by the client and the interface offered by the server. Note that in a response the roles of schema compliance are reversed (the server is the sender and the client is the receiver).

- § **Server-server: Conformance** (Carmona, van Dongen, Solti, & Weidlich, 2018), which expresses the replacement compatibility between two servers. The issue is to ascertain whether a server S_1 , serving a client, can be replaced by another server S_2 such that the client-server relationship enjoyed by the client is not impaired, or whether server S_2 is replacement compatible with server S_1 . Server S_2 must possess at least all the characteristics of server S_1 , therefore being able to take the form of (*conform to*) server S_1 and fulfill the expectations of the client regarding what it expects the server to be. In particular, the schema for request messages in server S_2 cannot include any additional mandatory component, regarding the schema of server S_1 , otherwise it would impose more requirements on the client and the interoperability could break.

Figure 2. Compliance and conformance



The server does not have to know the schema used by the client to generate messages, nor the client needs to know the full set of request messages accepted by the server. In this context, two kinds of schema are considered:

- § **Type schema:** A description of the features of a set of messages, including the definition of each possible feature and whether it is optional or mandatory. This is what typically schema description languages such as XML-Schema provide.
- § **Value schema:** The concrete definition of a specific message, with just the features that it actually has. Each message has its own value schema. This is simply a self-description and, unlike type

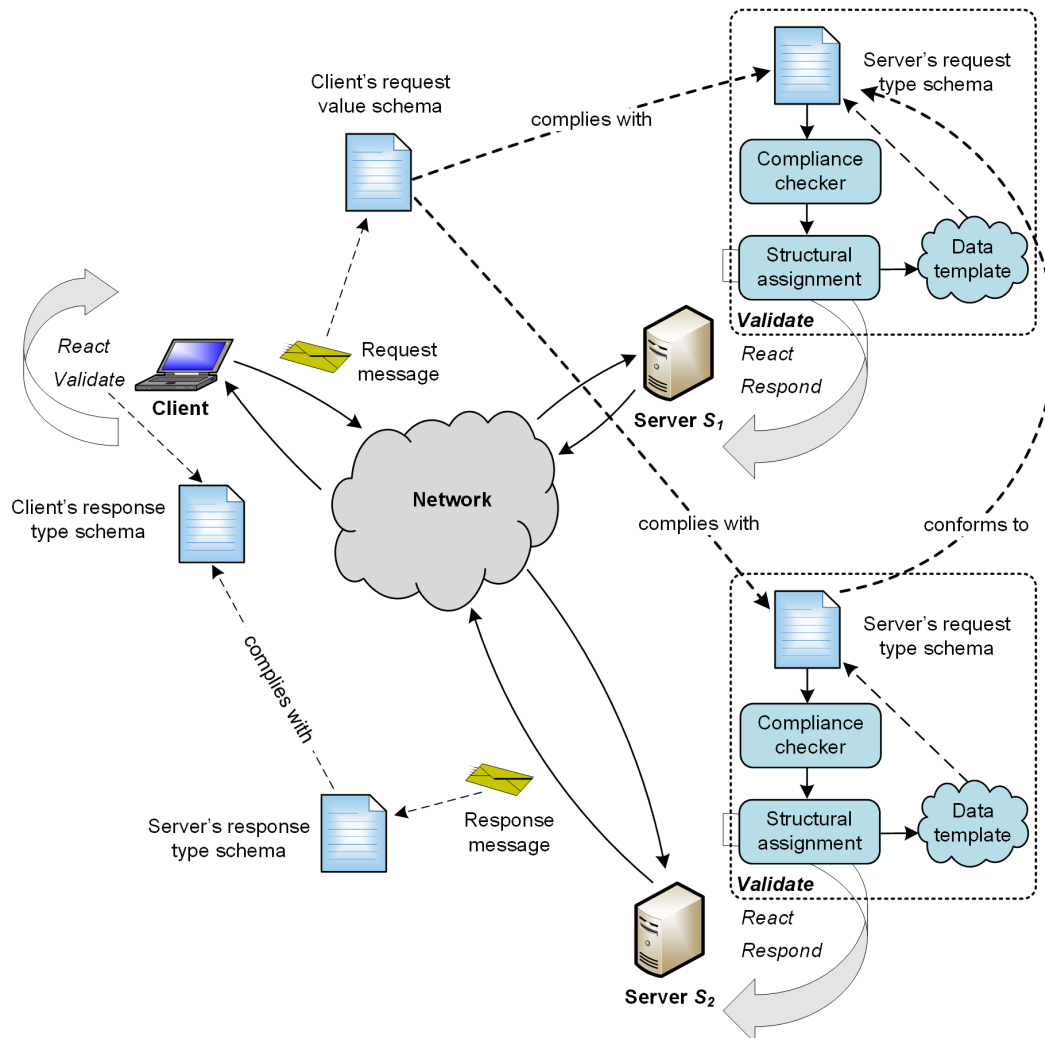
schemas, includes no variability (range of structured values). A value schema satisfies an infinite number of type schemas (by considering an infinite number of optional features).

Given that the value schemas of the messages actually sent by clients satisfy the type schema of the server, interoperability is possible, with a lower coupling than by sharing schemas. In semantic terms, compliance means that the set of possible message values sent by a client is a subset of the set of values that satisfy the type schema of the server. Conformance means that the set of values that satisfy the type schema of an alternative server (Figure 2) is a superset of the set of values that satisfy the type schema of the original receiver.

As long as compliance and conformance hold, the server can accept messages from different clients. In addition, a client can start using an alternative server without noticing the difference with respect to the original server.

Figure 3 illustrates both compliance and conformance. A client initially used server S_1 , under a compliance relationship. Then server S_2 , conforming to S_1 , replaced it. The client now complies with server S_2 and can use it as if it were server S_1 .

Figure 3. Illustrating compliance and conformance



The compliance-based interoperability mechanism can be detailed in the following way, for the request message (similar for the response message, with client and server roles reversed):

- § The server publishes a request type schema, which describes the type of request message values that the server can accept.
- § The actual request message, sent by the client, is described by a value schema (specific of this message).
- § When the request message arrives at the server, the message's value schema is validated by checking it for satisfaction of (compliance with) the type schema of the server, in the *compliance checker*. If compliance holds, the request message is one of those that satisfy the server's request type schema and is accepted.
- § The request message's value is *structurally* assigned to the *data template*, which is a data structure that satisfies the server's type schema and includes default values for the components in the type schema that are *optional* (minimum number of occurrences specified as zero). Compliance will fail if the request message does not include at least the minimum number of occurrences specified for each component of data template.
- § Structural assignment involves mapping the request message to the server's request type schema, by assigning the message to the data template, component by component, according to the following basic rules:
 - Components in the message that do not comply with any component in the data template are not assigned and simply ignored.
 - Components in the data template that are also present in the message have their values set to the corresponding message's component values.
 - Optional components in the data template missing from the message have their values set to the corresponding default values (specified in the data template).
 - Structured components are assigned by recursive application of these rules.

After this, the data template is completely populated and ready to be accessed by the server. Each request message populates a new instance of the data template. Note that this mechanism is different from the usual data binding of existing technologies, since the server deals only with its own request message type schema. It does not know the value schema of the request message and there is no need for a data binding stub to deal with it. The important issue is that the mapping between the request message and the data template is done in a universal manner (based on primitive data types and on structuring rules) and does not depend on the schemas used by either the client or the server. As long as compliance holds, the structural assignment rules can be applied. This means that coupling is reduced in comparison to classical schema-sharing technologies.

In conformance between the two servers, server S_2 has all the characteristics of server S_1 (and probably more) and is therefore able to take the form of (conform to) server S_1 and fulfill the expectations of the client. If the schema of server S_2 has additional features, with respect to server S_1 , these must be optional, otherwise it would impose more requirements on the client and the interoperability could be impaired.

By the definition of conformance, a client that complies with a server also complies with another server that conforms to the original server, as shown in Figure 3.

FUTURE RESEARCH DIRECTIONS

Compliance and conformance are basic concepts that can be applied to all domains and abstraction levels of Table 1. Although work exists on its formal treatment in specific areas, such as choreographies, an encompassing and systematic study needs to be conducted to formalize the definition of compliance and conformance.

The coupling problem is also relevant for the interoperability of non-functional aspects, namely in context-aware applications and in those involving the design and management of SLR (Service Level Requirements) with respect to enterprise-class applications and their supply chain interactions. Detailing how compliance and conformance can be applied in these cases requires additional research.

In addition, there is currently no practical implementation of an integration technology based on compliance and conformance. There are several possibilities on how to proceed to a practical implementation, namely:

- § In Web Services, to extend WSDL files with another section (which could be designated Structure), in which component resources and links can be exposed, thereby allowing Web Services to have structure.
- § To allow REST resources to have their own service description (e.g., an extended WADL file), thereby allowing the coexistence of the uniform and non-uniform interface approaches in REST.

CONCLUSION

Application integration is a fundamental problem in distributed information systems and entails two slants, interoperability and coupling. Current integration technologies, namely Web Services (Zimmermann, Tomlinson, & Peuser, 2012) and RESTful APIs (Pautasso, 2014), focus more on the interoperability side of the issue than on coupling. They require sharing of the data schemas by the interacting applications, which leads to more coupling than actually needed. Web Services usually require a link to the schema shared by the client and by the server, whereas RESTful applications usually agree beforehand on a given schema, or media type. If an application changes the schema it uses, the others interacting with it must adapt to these changes in order to maintain the interaction.

Both aspects, interoperability and coupling, need to be dealt with in a balanced way. The fundamental problem of application integration is how to achieve the minimum possible coupling while ensuring the minimum interoperability requirements, so that dependencies between applications that hinder changeability are kept to a minimum and applications are able to effectively interact.

Structural compliance (Czepa, Tran, Zdun, Kim, Weiss, & Ruhsam, 2017) and conformance (Carmona, van Dongen, Solti, & Weidlich, 2018) relax the constraint of having to share message schemas and therefore constitute an improved solution over existing integration technologies. Any two applications can interoperate as long as that share at least the characteristics actually required by interoperability, independently of knowing the rest of the characteristics, thereby minimizing coupling and increasing the number of servers that are compatible with a client, and the number of clients that are compatible with a server.

REFERENCES

- Agostinho, C., Ducq, Y., Zacharewicz, G., Sarraipa, J., Lampathaki, F., Poler, R., & Jardim-Goncalves, R. (2016). Towards a sustainable interoperability in networked enterprise information systems: Trends of knowledge and model-driven technology. *Computers in Industry*, 79, 64–76. doi:10.1016/j.comp-ind.2015.07.001
- Alenezi, M., & Magel, K. (2014). Empirical evaluation of a new coupling metric: Combining structural and semantic coupling. *International Journal of Computers and Applications*, 36(1). Advance online publication. doi:10.2316/Journal.202.2014.1.202-3902
- Babu, D., & Darsi, M. (2013). A Survey on Service Oriented Architecture and Metrics to Measure Coupling. *International Journal on Computer Science and Engineering*, 5(8), 726–733.
- Bassett, L. (2015). *Introduction to JavaScript Object Notation: A to-the-point Guide to JSON*. O'Reilly Media, Inc.
- Bidve, V. S., & Sarasu, P. (2016). Tool for measuring coupling in object-oriented java software. *IACSIT International Journal of Engineering and Technology*, 8(2), 812–820.
- Bora, A., & Bezboruah, T. (2015). A Comparative Investigation on Implementation of RESTful versus SOAP based Web Services. *International Journal of Database Theory and Application*, 8(3), 297–312. doi:10.14257/ijdta.2015.8.3.26
- Brandt, C., & Hermann, F. (2013). Conformance analysis of organizational models: A new enterprise modeling framework using algebraic graph transformation. *International Journal of Information System Modeling and Design*, 4(1), 42–78. doi:10.4018/jismd.2013010103
- Capel, M., & Mendoza, L. (2014). Choreography Modeling Compliance for Timed Business Models. In *Proceedings of the Workshop on Enterprise and Organizational Modeling and Simulation* (pp. 202–218). Berlin, Germany: Springer. 10.1007/978-3-662-44860-1_12
- Carmona, J., van Dongen, B., Solti, A., & Weidlich, M. (Eds.). (2018). *Conformance Checking: Relating Processes and Models*. Springer. doi:10.1007/978-3-319-99414-7
- Chinosi, M., & Trombetta, A. (2012). BPMN: An introduction to the standard. *Computer Standards & Interfaces*, 34(1), 124–134. doi:10.1016/j.csi.2011.06.002
- Czepa, C., Tran, H., Zdun, U., Kim, T., Weiss, E., & Ruhsam, C. (2017). On the understandability of semantic constraints for behavioral software architecture compliance: A controlled experiment. In *Proceedings of the International Conference on Software Architecture* (pp. 155–164). Piscataway, NJ: IEEE Computer Society Press. 10.1109/ICSA.2017.10
- Dahiya, N., & Parmar, N. (2014). SOA AND REST Synergistic Approach. *International Journal of Computer Science and Information Technologies*, 5(6), 7045–7049.
- Dornberger, R. (Ed.). (2018). *Business Information Systems and Technology 4.0: New Trends in the Age of Digital Change* (Vol. 141). Springer. doi:10.1007/978-3-319-74322-6
- Erl, T., Merson, P., & Stoffers, R. (2017). *Service-oriented Architecture: Analysis and Design for Services and Microservices*. Prentice Hall PTR.

- Fawcett, J., Ayers, D., & Quin, L. (2012). *Beginning XML*. John Wiley & Sons.
- Fielding, R., Taylor, R., Erenkrantz, J., Gorlick, M., Whitehead, J., Khare, R., & Oreizy, P. (2017). Reflections on the REST architectural style and principled design of the modern web architecture. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* (pp. 4-14). New York, NY: ACM Press. 10.1145/3106237.3121282
- Graydon, P., Habli, I., Hawkins, R., Kelly, T., & Knight, J. (2012). Arguing Conformance. *IEEE Software*, 29(3), 50–57. doi:10.1109/MS.2012.26
- Hendricksen, D. (2014). *12 More Essential Skills for Software Architects*. Addison-Wesley Professional.
- Henning, M. (2008). The rise and fall of CORBA. *Communications of the ACM*, 51(8), 52–57. doi:10.1145/1378704.1378718
- ISO. (2010). *Systems and software engineering – Vocabulary. ISO/IEC/IEEE 24765:2010(E) International Standard*. International Organization for Standardization.
- Kukreja, R., & Garg, N. (2014). Remote Procedure Call: Limitations and Drawbacks. *International Journal of Research*, 1(10), 914–917.
- Liao, Y., Deschamps, F., Loures, E., & Ramos, L. (2017). Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal. *International Journal of Production Research*, 55(12), 3609–3629. doi:10.1080/00207543.2017.1308576
- Page, T., & Thorsteinsson, G. (2018). Emerging Technologies in Interaction with Mobile Computing Devices-A Technology Forecast. *Journal on Mobile Applications and Technologies*, 5(1), 1. doi:10.26634/jmt.5.1.14694
- Panetto, H., & Whitman, L. (2016). Knowledge engineering for enterprise integration, interoperability and networking: Theory and applications. *Data & Knowledge Engineering*, 105, 1–4. doi:10.1016/j.datak.2016.05.001
- Paul, P., & Saraswathi, R. (2017). The Internet of Things – A comprehensive survey. In *Proceedings of the International Conference on Computation of Power, Energy Information and Communication* (pp. 421-426). Piscataway, NJ: IEEE Computer Society Press. 10.1109/ICCPEIC.2017.8290405
- Pautasso, C. (2014). RESTful web services: principles, patterns, emerging technologies. In *Web Services Foundations* (pp. 31-51). New York, NY: Springer. doi:10.1007/978-1-4614-7518-7_2
- Pautasso, C., Zimmermann, O., & Leymann, F. (2008). Restful web services vs. “big” web services: making the right architectural decision. In *International conference on World Wide Web* (pp. 805-814). ACM Press. 10.1145/1367497.1367606
- Preidel, C., & Borrmann, A. (2016). Towards code compliance checking on the basis of a visual programming language. *Journal of Information Technology in Construction*, 21(25), 402–421.
- Sikos, L. (2017). The Semantic Gap. In *Description Logics in Multimedia Reasoning* (pp. 51–66). Springer. doi:10.1007/978-3-319-54066-5_3

Sungkur, R., & Daiboo, S. (2016). Combining the Best Features of SOAP and REST for the Implementation of Web Services. *International Journal of Digital Information and Wireless Communications*, 6(1), 21–33. doi:10.17781/P001923

Varghese, B., & Buyya, R. (2018). Next generation cloud computing: New trends and research directions. *Future Generation Computer Systems*, 79, 849–861. doi:10.1016/j.future.2017.09.020

Verborgh, R., Harth, A., Maleshkova, M., Stadtmüller, S., Steiner, T., Taheriyan, M., & Van de Walle, R. (2014). Survey of semantic description of REST APIs. In C. Pautasso, E. Wilde, & R. Alarcon (Eds.), *REST: Advanced Research Topics and Practical Applications* (pp. 69–89). Springer. doi:10.1007/978-1-4614-9299-3_5

Wang, H., Gibbins, N., Payne, T., Patelli, A., & Wang, Y. (2015). A survey of semantic web services formalisms. *Concurrency and Computation*, 27(15), 4053–4072. doi:10.1002/cpe.3481

Zimmermann, O., Tomlinson, M., & Peuser, S. (2012). *Perspectives on Web Services: Applying SOAP, WSDL and UDDI to Real-World Projects*. Springer Science & Business Media.

ADDITIONAL READING

Erl, T., Carlyle, B., Pautasso, C., & Balasubramanian, R. (2012). *SOA with REST: Principles, Patterns & Constraints for Building Enterprise Solutions with REST*. Prentice Hall Press.

Geetika, R., & Singh, P. (2014). Dynamic coupling metrics for object oriented software systems: A survey. *Software Engineering Notes*, 39(2), 1–8. doi:10.1145/2579281.2579296

He, W., & Da Xu, L. (2014). Integration of distributed enterprise applications: A survey. *IEEE Transactions on Industrial Informatics*, 10(1), 35–42. doi:10.1109/TII.2012.2189221

Kostoska, M., Gusev, M., & Ristov, S. (2016). An overview of cloud interoperability. In *Federated Conference on Computer Science and Information Systems* (pp. 873–876). Piscataway, NJ: IEEE Computer Society Press. 10.15439/2016F463

Mezgár, I., & Rauschecker, U. (2014). The challenge of networked enterprises for cloud computing interoperability. *Computers in Industry*, 65(4), 657–674. doi:10.1016/j.compind.2014.01.017

Ritter, D., May, N., & Rinderle-Ma, S. (2017). Patterns for emerging application integration scenarios: A survey. *Information Systems*, 67, 36–57. doi:10.1016/j.is.2017.03.003

Sharma, A., Kumar, M., & Agarwal, S. (2015). A Complete Survey on Software Architectural Styles and Patterns. *Procedia Computer Science*, 70, 16–28. doi:10.1016/j.procs.2015.10.019

Thakar, U., Tiwari, A., & Varma, S. (2016). On Composition of SOAP Based and RESTful Services. In *Proceedings of the 6th International Conference on Advanced Computing* (pp. 500–505). Piscataway, NJ: IEEE Computer Society Press. 10.1109/IACC.2016.99

KEY TERMS AND DEFINITIONS

Client: A role performed by a resource C in an interaction with another S , which involves making a request to S and typically waiting for a response.

Compliance: Asymmetric property between a client C and a server S (C is compliant with S) that indicates that C satisfies all the requirements of S in terms of accepting requests.

Conformance: Asymmetric property between two servers, S_1 and S_2 (S_1 conforms to S_2) that indicates that S_1 fulfills all the expectations of all the clients of S_2 in terms of the effects caused by its requests, which means that S_2 can replace S_1 .

Coupling: A measurement of how much an application is dependent on the interface of another application.

Interoperability: Asymmetric property between a client C and a server S that holds if C is compliant with the set of requests allowed by S and S is conformant with the set of responses expected by C .

Resource: The computer-based implementation of a service.

Server: A role performed by a resource S in an interaction with another C , which involves waiting for a request from C , honoring it and typically sending a response back to C .

Service: The set of operations supported by an application that together define its interface (the set of reactions to messages that the application is able to receive and process) and behavior.

A Survey of Tasks Scheduling Algorithms in Distributed Computing Systems

2

Nutan Kumari Chauhan

KIET Group of Institutions, India

Harendra Kumar

 <https://orcid.org/0000-0002-3394-5655>

Gurukula Kangri Vishwavidyalaya, India

INTRODUCTION

There are various causes for using DCS. The nature of equipment may involve the utilization of a communication network which connected by some computers: for eg, data created in one site and needed in another site. There are various cases in which only one computer is required, but DCS is very helpful for practical causes. For eg., it may be extra cost-efficient to get the inclined level of performance by applying cluster of numerous low-end computers, in similarity with only high-end computer. DCS has no single point of failure, so it can offer additional reliability than a non-distributed system. DCS may be simple to expand and run than a uni-processor system. DCS have used in a large scale due to several characteristics such as communication, computational speedup, fault tolerance, increased throughput, resource sharing etc.

Tasks scheduling is very important phase in DCS, mainly it is assigning of tasks onto processors as a manner so that some efficiency actions are optimized. An effective tasks scheduling policy avoid too much inter-tasks communication and utilize the precise effectiveness of the processors. It is extremely important step in exploiting the capacity of system and it may be done in two ways: **Dynamic Scheduling** and **Static Scheduling**. The most important plan of tasks scheduling is toward decrease the running cost of a program. The cost can be characterized by time, money or some other measure of resources usage.

In DCS, resources of system can be split by users at different sites through the communication link. However in proper utilization of total computing power of DCS, a fundamental problem arises. It may happen that a certain processor has very few workloads while another processor has more workload at a given time. It is required to distribute total workload of DCS of its overall processors. This evades under use of power and further it decreases response time for deputize at more heavily loaded processor. Generally, the form of computing power sharing with the intention of getting better performance of a DCS by rearranging the workload between the available processors called load balancing. The progress in the presentation of system through redistributing the workload among processors is the main aim of load balancing policies.

The reliability of DCS is probability that system can run the whole application successfully. If system is complex then it is very difficult to system in it's entirely. The logical technique is applied to crumble the entire system into function entities compiled of subsystems and units. Each entity is either operational or failed. The sub-division generates the block diagram of system. Models are prepared to fit the logical arrangement and calculus of probability is applied to figure out reliability of system in terms of sub-division reliabilities. There are two types of reliabilities in DCS: first is processor-related reliability,

DOI: 10.4018/978-1-7998-3479-3.ch018

Table 1. Difference between serial and parallel processing

S. No.	Serial Processing	Parallel Processing
1.	There exist several tasks and processors even although only single task is active on a processor at a given time	More than one task of the program may execute concurrently for various periods during the lifetime of a program
2.	The main concern in a serial processing of programs is variability or non-uniformity of computation requirements and the overhead of communication	The main concern of parallel processing is to reduce the total time of the program by running different parts of the program
3.	The challenge of distributed processing lies in matching requirements against resources without incurring excessive communication overhead	In parallel processing, arranging the order of execution of the various tasks on the processors is a major problem
4.	Generally, serial processing is used for long distance	Generally, parallel processing is used for short distance

for example reliability of resources and reliability of computation; another is the path / link-related reliability, for example reliability of communication paths between allocated processors.

The processing of a program is of two types: serial processing or parallel processing.

Difference between serial processing and parallel processing are given in the following table-1:

Figure 1 and Figure 2 show the execution of tasks on processors in a serial and parallel way respectively. A serial program made up of five tasks is executing on a single processor in Figure 1(a), while the same serial program is executed over a two processors system in Figure 1(b). Processor P2 can execute two tasks (3 and 4) more efficiently but due to transfers of control between tasks over a communication link incur appreciable overhead.

Figure 1. Serial Processing

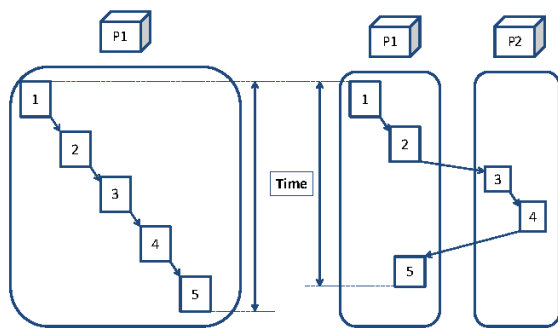
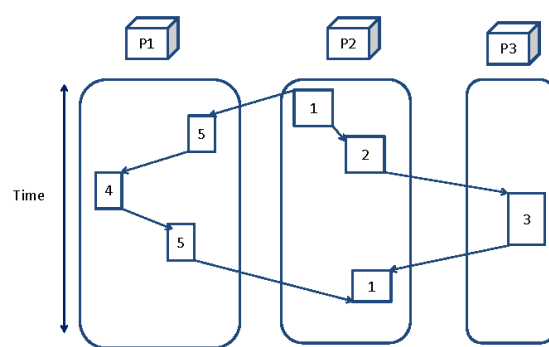


Figure 2. Parallel Processing

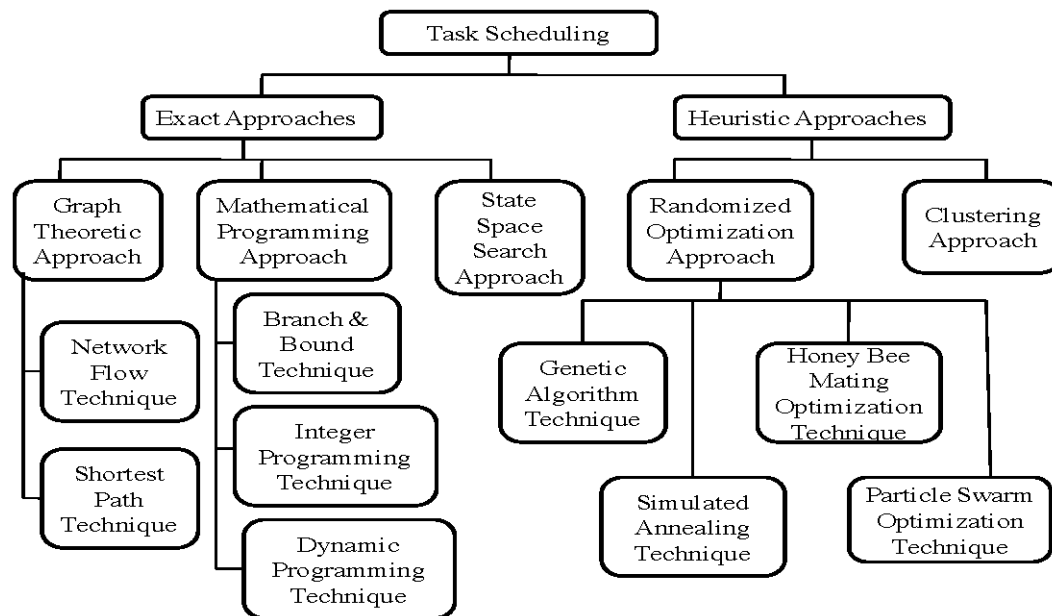


Further the chapter is planned as follows: Section 2 portrays the background of task scheduling techniques. Section 3 covers the application of task scheduling in DCS. Section 4 describes the future research directions. In section 5, conclusion is discussed on the basis of various factors in task scheduling techniques.

BACKGROUND

In this section different approaches for solving tasks scheduling problem with one or more objectives are discussed. These approaches are characterized according to the taxonomy given in the following Figure 3.

Figure 3. Classification of Tasks Scheduling Approach



Exact Approaches

Exact approaches yield optimal solution in a period by limiting the formation of program. Exact approaches are classified into the following approaches:

- Graph Theoretic Approach
- Mathematical Programming Approach
- State Space Search Approach

Graph Theoretic Approach

It is one of the important approaches for optimization in DCS. It uses the graphical method to represents and allocation of tasks onto the processors. Further it may be classified into two techniques: **Network Flow Technique** and **Shortest Path Technique**

- **Network Flow Technique**

A directed graph in which each edge has a capacity and gets a flow is called network flow or transportation network. The total of flow on an edge cannot beat the capacity of the edge. [(Stone, 1977) and (Lee & Shin, 1997)] uses network flow algorithm for tasks scheduling in DCS.

- **Shortest Path Technique**

In shortest path technique, Program chart is changed into assignment chart to apply shortest path technique; it assesses set of processors in assignment chart to select most excellent possible way for scheduling of tasks to the processors. The processors and edges among them are tagged accordingly on method used to get shortest path from source-processor to terminal-processor. [(Bokhari, 1981), (Hui & Chanson, 1997), (Ying, Shakkottai, & Reddy, 2009) and (Meng, Li, Liu, & Ngan, 2012)] used shortest path scheme to solve tasks scheduling problem with purpose of minimization of system cost.

Mathematical Programming Approach

It is used at a large scale for modelling and analyzing several kinds of problems. Initially it converts the tasks scheduling problem into optimization problem and then solve. The following techniques are included in this approach:

- Branch & Bound Technique
- Integer Programming Technique
- Dynamic Programming Technique
- Branch & Bound Technique

It is extremely popular and widely accepted techniques by researchers. [(Kartik & Murthy, 1995), (Kim, Lee, & Lee, 2015), and (Lu, Zhou, Yue, & Chen, 2016)] discussed the Branch & Bound Technique for scheduling the tasks in DCS.

- **Integer Programming Technique**

It is mathematical optimization, in which few or all variables are limited to be an integer. A tasks scheduling model becomes additional noteworthy and sensible when it adds real-time constraints for instance execution, memory limits of each processor, inter-tasks communication, etc. Integer programming is very useful and exhaustive approach for tasks scheduling problems as it is able to replicate real-life conditions of distributed processing and easy in an application. Several researchers worked on tasks scheduling problems with different objective by using this technique. [(Manić, Bahiense, & de Souza, 2009) and (Jabr, Singh, & Pal, 2012)] proposed integer programming formulation for the least failure distribution system reconfiguration problem.

- **Dynamic Programming Technique**

Dynamic programming technique is applied for both mathematical optimization and computer programming. It is an algorithm formation technique which can be utilized when outcome of problem

may be observed as outcome of sequence of result. Various research papers [(Bokhari, 1981),(Ashrafi & Berman, 1992), (Dhakal, Hayat, Pezoa, Yang, & Bader, 2007), (Mishra, Mishra, Mishra, & Tripathi, 2012), (H. Kumar, Chauhan, & Yadav, 2016), (Janati et al., 2017)] use dynamic programming technique to get the result of tasks scheduling problem in DCS.

State Space Search Approach

To make the best use of this approach for tasks scheduling problem, first the problem is switched into a state space search tree and then cost function is defined for leading search. The path defined by the chain of operators which show the ways an initial processor to one of the external processor is the solution path. [(Shen & Tsai, 1985), (Shatz, Wang, & Goto, 1992),(Doğan & Özgüner, 2000) and (Sekhar, Manoj, & Siva Ram Murthy, 2005)] worked on A* algorithm to discover the solution of tasks scheduling problem for the optimization of system cost and system reliability.

Heuristic Approaches

Although an optimal solution of the tasks scheduling problem can be find by using exact approach but generally, it is a NP-hard problem. Therefore finding optimal solution with exact approach to large scale problem is computationally prohibitive. Hence, heuristic approaches play an important task for offering solutions to different kinds of problems. Heuristic approaches provide near to optimal solution for tasks scheduling problem. (Arora & Rana, 1980) used heuristic approaches for tasks scheduling problems in a DCS. Further Heuristic approaches can be characterized into two types: Randomized Optimization Approach and Clustering Approach

Randomized Optimization Approach

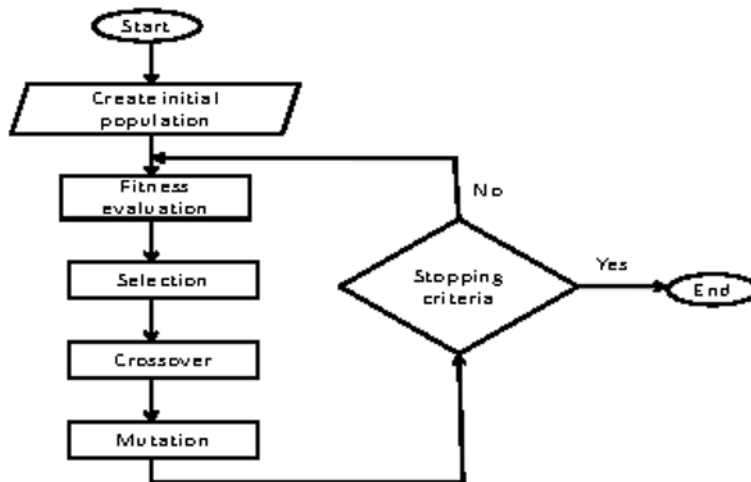
It is a probabilistic approach, which applied to discover the solution of tasks scheduling problem. It begins from as an initial solution and searches for the progress in the scheduling by switching and shifting tasks between processors. From this approach, the following techniques are derived to obtain the solution of tasks scheduling problem:

- Genetic Algorithm Technique
- Honey bee Optimization Technique
- Particle Swarm Optimization Technique
- Simulated Annealing Technique
- Genetic Algorithm Technique

It is foundation upon the Darwin's theory of "survival of fittest". It follows the behavior of the reproduction in nature. The main steps of genetic algorithm are reproduction, selection, crossover and mutation. Selection, mutation and crossover and are called genetic operators and used repeatedly until the termination condition is not satisfied. The block diagram of genetic algorithm is given in Figure 4.

[(I. Ahmad, Dhodhi, & Ghafoor, 1995), (Mishra et al., 2012),(Tripathi, 2000), (Hsieh & Hsieh, 2003) and (Yang, Hu, & Guo, 2009), (Omara & Arafa, 2009), (S. G. Ahmad, Liew, Munir, Ang, & Khan, 2016), (H. Kumar, Chauhan, & Yadav, 2019)] proposed genetic algorithm for tasks scheduling in DCS. It is also used for load balancing through the execution face to make use of resources at maximum level.

Figure 4. Block Diagram of Genetic Algorithm



- Honey Bee Mating Optimization Technique**

It is based on meta-heuristic which motivated by the public group of honeybees and their process of reproduction. In this technique, simulated annealing coincides to mating procedure of honeybees; genetic algorithm replicates the breeding process and application based local search replicates the process of nourishing the children and queen bees. [(Abbass & Abbass, 2001), (Fathian, Amiri, & Maroosi, 2007), (Q. M. Kang, He, Song, & Deng, 2010), (Q. Kang & He, 2013) and (Jin, Zhang, & Shao, 2015) build up a hybrid honey bee mating optimization algorithm to finding the solution of tasks scheduling problem with the optimization of system cost.

- Particle Swarm Optimization Technique**

It is a meta-heuristic optimization method and introduced by Kennedy and Eberhart (1995) and induced by the actions of fishes schooling and bird flocking. It has some principle mechanisms such as: particle representation, swarm, experience and stopping criterion. It is very similar to genetic algorithm. In this technique, each swarm of individuals in search space is a candidate result to the optimization problem. Each element shifts in the region of in multi-dimensional search space with dynamically velocity and regulated according to the flying skill of the individuals and its generation. This technique increases the stochastic behavior of the particle and meets quickly to global minima near to optimal solution. It has a great application to solve tasks scheduling problem in DCS. [(Salman, Ahmad, & Al-Madani, 2002), (Yin, Yu, Wang, & Wang, 2006), (Zhang, Chen, Sun, & Yang, 2008), (Q. Kang & He, 2011) and (N. Kumar & Vidyarthi, 2016)] uses particle swarm technique for tasks scheduling in DCS.

- Simulated Annealing Technique** It is a global optimization technique and applied to get the lowest peak in energy land space. It follows the physical conceptions of temperature and power of dynamic system while temperature is launched to randomize the find a solution. Several researchers [(Rutenbar, 1989), (Attiya & Hamam, 2006), (He, Gu, & Zhu, 2010), (Damodaran & Vélez-Gallego, 2012), (Arya & Dhingra, 2015), (X. Liu & Liu, 2016) and (Wu, Wang, Pedrycz, Li, & Wang, 2017)] have used simulated annealing approach to get the solution of tasks scheduling problem in DCS.

2.2.2 Clustering Approach

Mostly research work on heuristic algorithms has dedicated to minimization of inter-tasks communication cost. Clustering approach is one of the heuristic approaches which is applied to decrease the inter-task communication cost by clustering the highly communicated tasks together. Generally, the number of tasks clusters and processors must be equal to reduce the size of program in system so that it could be extremely supportive to get a good quality scheduling. [(Sarje & Sagar, 1991), (Abdelzaher & Shin, 2000), (Lin, Lin, Lin, Hsiung, & Shih, 2013), (Li, Tang, Veeravalli, & Li, 2015), (Y. Liu, Zhang, Li, & Niu, 2017) and (Singh, Gupta, & Jana, 2018), (H. Kumar, Chauhan, & Yadav, 2018)] used clustering approach to solve tasks scheduling problem in DCS.

3. Applications of DCS

DCS has widely used in a lot of fields such as:

- Telephone networks.
- Nuclear Reactor.
- Network files systems.
- Computer networks such as internet.
- Aircraft control system.
- Industrial control system.
- Distributed rendering in computer graphics.
- Wireless sensor networks

FUTURE RESEARCH DIRECTION

Tasks scheduling in DCS is a wide area for research. The effort of categorizing tasks scheduling does not come to a close to end. Several researchers have applied various techniques for solving the tasks scheduling problem such as: genetic algorithm, honey bee mating optimization, particle swarm optimization etc. Use of soft computing techniques such as neural network and various modified soft clustering technique can also be used for tasks scheduling in future. The need for classification and survey of tasks scheduling literature for DCS will remain in the future. A very few research has been carried out on dynamic tasks scheduling and load balancing problems. Therefore, for the ample use of DCS, dynamic tasks scheduling and load balancing are an important issue for further investigation.

CONCLUSION

The classification of tasks scheduling techniques is very demanding due to the huge amount of literature as well as the number of variations of the objectives of the problem. In this chapter we review numerous aspects of the tasks scheduling literature for distributed computing systems, and proposed taxonomy. To confirm the taxonomy, we reviewed various research papers from the scheduling literature, identifying research according to proposed taxonomy. Each of them performs well on the basis of some performance measures. Each of them is compared on the origin of various performance evaluation factors like

computational cost, execution cost, total cost, response time and system reliability etc. This resulted in a study of the latest scheduling research in DCS from variety of features.

REFERENCES

- Abbass, H., & Abbass, H. (2001). A Monogenous MBO Approach to Satisfiability. *International conference on computational intelligence for modelling, control and automation cimca'2001*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.9327>
- Abdelzaher, T. F., & Shin, K. G. (2000). Period-based load partitioning and assignment for large real-time applications. *IEEE Transactions on Computers*, 49(1), 81–87. doi:10.1109/12.822566
- Ahmad, I., Dhodhi, M. K., & Ghafoor, A. (1995). Task assignment in distributed computing systems. *Conference Proceedings - International Phoenix Conference on Computers and Communications*, 49–53. 10.1109/PCCC.1995.472512
- Ahmad, S. G., Liew, C. S., Munir, E. U., Ang, T. F., & Khan, S. U. (2016). A hybrid genetic algorithm for optimization of scheduling workflow applications in heterogeneous computing systems. *Journal of Parallel and Distributed Computing*, 87, 80–90. doi:10.1016/j.jpdc.2015.10.001
- Arora, R. K., & Rana, S. P. (1980). Heuristic algorithms for process assignment in distributed computing systems. *Information Processing Letters*, 11(4–5), 199–203. doi:10.1016/0020-0190(80)90050-2
- Arya, S., & Dhingra, S. (2015). *Analysis of Heuristic Based Hybrid Simulated Annealing Algorithm for Multiprocessor Task Scheduling*. doi:10.5281/ZENODO.1110229
- Ashrafi, N., & Berman, O. (1992). Optimization Models for Selection of Programs, Considering Cost & Reliability. *IEEE Transactions on Reliability*, 41(2), 281–287. doi:10.1109/24.257794
- Attiya, G., & Hamam, Y. (2006). Task allocation for maximizing reliability of distributed systems: A simulated annealing approach. *Journal of Parallel and Distributed Computing*, 66(10), 1259–1266. doi:10.1016/j.jpdc.2006.06.006
- Bokhari, S. H. (1981). A Shortest Tree Algorithm for Optimal Assignments Across Space and Time in a Distributed Processor System. *IEEE Transactions on Software Engineering*, SE, 7(6), 583–589. doi:10.1109/TSE.1981.226469
- Damodaran, P., & Vélez-Gallego, M. C. (2012). A simulated annealing algorithm to minimize makespan of parallel batch processing machines with unequal job ready times. *Expert Systems with Applications*, 39(1), 1451–1458. doi:10.1016/j.eswa.2011.08.029
- Dhakal, S., Hayat, M. M., Pezoa, J. E., Yang, C., & Bader, D. A. (2007). Dynamic load balancing in distributed systems in the presence of delays: A regeneration-theory approach. *IEEE Transactions on Parallel and Distributed Systems*, 18(4), 485–497. doi:10.1109/TPDS.2007.1009
- Doğan, A., & Özgüner, F. (2000). Reliable matching and scheduling of precedence-constrained tasks in heterogeneous distributed computing. *Proceedings of the International Conference on Parallel Processing*, 307–314. 10.1109/ICPP.2000.876146

- Fathian, M., Amiri, B., & Maroosi, A. (2007). Application of honey-bee mating optimization algorithm on clustering. *Applied Mathematics and Computation*, 190(2), 1502–1513. doi:10.1016/j.amc.2007.02.029
- He, X., Gu, Z., & Zhu, Y. (2010). Task allocation and optimization of distributed embedded systems with simulated annealing and geometric programming. *The Computer Journal*, 53(7), 1071–1091. doi:10.1093/comjnl/bxp084
- Hsieh, C. C., & Hsieh, Y. C. (2003). Reliability and cost optimization in distributed computing systems. *Computers & Operations Research*, 30(8), 1103–1119. doi:10.1016/S0305-0548(02)00058-8
- Hui, C. C., & Chanson, S. T. (1997). Allocating task interaction graphs to processors in heterogeneous networks. *IEEE Transactions on Parallel and Distributed Systems*, 8(9), 908–925. doi:10.1109/71.615437
- Jabr, R. A., Singh, R., & Pal, B. C. (2012). Minimum loss network reconfiguration using mixed-integer convex programming. *IEEE Transactions on Power Systems*, 27(2), 1106–1115. doi:10.1109/TPWRS.2011.2180406
- Janati, F., Abdollahi, F., Ghidary, S. S., Jannatifar, M., Baltes, J., & Sadeghnejad, S. (2017). *Multi-robot Task Allocation Using Clustering Method*. doi:10.1007/978-3-319-31293-4_19
- Jin, L., Zhang, C., & Shao, X. (2015). An effective hybrid honey bee mating optimization algorithm for integrated process planning and scheduling problems. *International Journal of Advanced Manufacturing Technology*, 80(5–8), 1253–1264. doi:10.100700170-015-7069-3
- Kang, Q., & He, H. (2011). A novel discrete particle swarm optimization algorithm for meta-task assignment in heterogeneous computing systems. *Microprocessors and Microsystems*, 35(1), 10–17. doi:10.1016/j.micpro.2010.11.001
- Kang, Q., & He, H. (2013). Honeybee Mating Optimization Algorithm For Task Assignment In Heterogeneous Computing Systems. *Intelligent Automation and Soft Computing*, 19(1), 69–84. doi:10.1080/10798587.2013.771438
- Kang, Q. M., He, H., Song, H. M., & Deng, R. (2010). Task allocation for maximizing reliability of distributed computing systems using honeybee mating optimization. *Journal of Systems and Software*, 83(11), 2165–2174. doi:10.1016/j.jss.2010.06.024
- Kartik, S., & Murthy, C. S. R. (1995). Improved Task-Allocation Algorithms to Maximize Reliability of Redundant Distributed Computing Systems. *IEEE Transactions on Reliability*, 44(4), 575–586. doi:10.1109/24.475976
- Kim, H. J., Lee, J. H., & Lee, T. E. (2015). Noncyclic Scheduling of Cluster Tools With a Branch and Bound Algorithm. *IEEE Transactions on Automation Science and Engineering*, 12(2), 690–700. doi:10.1109/TASE.2013.2293552
- Kumar, H., Chauhan, N. K., & Yadav, P. K. (2016). Dynamic tasks scheduling algorithm for distributed computing systems under fuzzy environment. *International Journal of Fuzzy System Applications*, 5(4), 77–95. doi:10.4018/IJFSA.2016100104
- Kumar, H., Chauhan, N. K., & Yadav, P. K. (2018). A high performance model for task allocation in distributed computing system using k-means clustering technique. *International Journal of Distributed Systems and Technologies*, 9(3), 1–23. doi:10.4018/IJDST.2018070101

- Kumar, H., Chauhan, N. K., & Yadav, P. K. (2019). Hybrid genetic algorithm for task scheduling in distributed real-time system. *International Journal of Systems. Control and Communications*, 10(1), 32–51. doi:10.1504/IJSCC.2019.097417
- Kumar, N., & Vidyarthi, D. P. (2016). A novel hybrid PSO–GA meta-heuristic for scheduling of DAG with communication on multiprocessor systems. *Engineering with Computers*, 32(1), 35–47. doi:10.1007/00366-015-0396-z
- Lee, C. H., & Shin, K. G. (1997). Optimal task assignment in homogeneous networks. *IEEE Transactions on Parallel and Distributed Systems*, 8(2), 119–129. doi:10.1109/71.577254
- Li, K., Tang, X., Veeravalli, B., & Li, K. (2015). Scheduling precedence constrained stochastic tasks on heterogeneous cluster systems. *IEEE Transactions on Computers*, 64(1), 191–204. doi:10.1109/TC.2013.205
- Lin, C. S., Lin, C. S., Lin, Y. S., Hsiung, P. A., & Shih, C. (2013). Multi-objective exploitation of pipeline parallelism using clustering, replication and duplication in embedded multi-core systems. *Journal of Systems Architecture*, 59(10), 1083–1094. doi:10.1016/j.sysarc.2013.05.024
- Liu, X., & Liu, J. (2016). A Task Scheduling Based on Simulated Annealing Algorithm in Cloud Computing. *International Journal of Hybrid Information Technology*, 9(6), 403–412. doi:10.14257/ijhit.2016.9.6.36
- Liu, Y., Zhang, C., Li, B., & Niu, J. (2017). DeMS: A hybrid scheme of task scheduling and load balancing in computing clusters. *Journal of Network and Computer Applications*, 83, 213–220. doi:10.1016/j.jnca.2015.04.017
- Lu, C., Zhou, L. S., Yue, Y. X., & Chen, R. (2016). A Branch and Bound Algorithm for the Exact Solution of the Problem of EMU Circulation Scheduling in Railway Network. *Mathematical Problems in Engineering*, 2016, 1–14. Advance online publication. doi:10.1155/2016/8537089
- Manić, G., Bahiense, L., & de Souza, C. (2009). A branch&cut algorithm for the maximum common edge subgraph problem. *Electronic Notes in Discrete Mathematics*, 35(C), 47–52. doi:10.1016/j.endm.2009.11.009
- Meng, F., Li, H., Liu, G., & Ngan, K. N. (2012). Object co-segmentation based on shortest path algorithm and saliency model. *IEEE Transactions on Multimedia*, 14(5), 1429–1441. doi:10.1109/TMM.2012.2197741
- Mishra, P. K., Mishra, A., Mishra, K. S., & Tripathi, A. K. (2012). Benchmarking the clustering algorithms for multiprocessor environments using dynamic priority of modules. *Applied Mathematical Modelling*, 36(12), 6243–6263. doi:10.1016/j.apm.2012.02.011
- Omara, F. A., & Arafa, M. M. (2009). Genetic algorithms for task scheduling problem. *Studies in Computational Intelligence*, 203, 479–507. doi:10.1007/978-3-642-01085-9_16
- Rutenbar, R. A. (1989). Simulated annealing algorithms: An overview. *IEEE Circuits and Devices Magazine*, 5(1), 19–26. doi:10.1109/101.17235
- Salman, A., Ahmad, I., & Al-Madani, S. (2002). Particle swarm optimization for task assignment problem. *Microprocessors and Microsystems*, 26(8), 363–371. doi:10.1016/S0141-9331(02)00053-4

Sarje, A. K., & Sagar, G. (1991). Heuristic model for task allocation in distributed computer systems. *IEE Proceedings. Computers and Digital Techniques*, 138(5), 313. doi:10.1049/ip-e.1991.0043

Sekhar, A., Manoj, B. S., & Siva Ram Murthy, C. (2005). A state-space search approach for optimizing reliability and cost of execution in distributed sensor networks. *Lecture Notes in Computer Science*, 3741, 63–74. doi:10.1007/11603771_7

Shatz, S. M., Wang, J. P., & Goto, M. (1992). Task Allocation for Maximizing Reliability of Distributed Computer Systems. *IEEE Transactions on Computers*, 41(9), 1156–1168. doi:10.1109/12.165396

Shen, C. C., & Tsai, W. H. (1985). A Graph Matching Approach to Optimal Task Assignment in Distributed Computing Systems Using a Minimax Criterion. *IEEE Transactions on Computers*, C-34(3), 197–203. doi:10.1109/TC.1985.1676563

Singh, V., Gupta, I., & Jana, P. K. (2018). A novel cost-efficient approach for deadline-constrained workflow scheduling by dynamic provisioning of resources. *Future Generation Computer Systems*, 79, 95–110. doi:10.1016/j.future.2017.09.054

Stone, H. S. (1977). Multiprocessor Scheduling with the Aid of Network Flow Algorithms. *IEEE Transactions on Software Engineering*, SE, 3(1), 85–93. doi:10.1109/TSE.1977.233840

Tripathi, A. K. (2000). A ga based multiple task allocation considering load. *International Journal of High Speed Computing*, 11(4), 203–214. doi:10.1142/S0129053300000187

Wu, G., Wang, H., Pedrycz, W., Li, H., & Wang, L. (2017). Satellite observation scheduling with a novel adaptive simulated annealing algorithm and a dynamic task clustering strategy. *Computers & Industrial Engineering*, 113, 576–588. doi:10.1016/j.cie.2017.09.050

Yang, B., Hu, H., & Guo, S. (2009). Cost-oriented task allocation and hardware redundancy policies in heterogeneous distributed computing systems considering software reliability. *Computers & Industrial Engineering*, 56(4), 1687–1696. doi:10.1016/j.cie.2008.11.001

Yin, P. Y., Yu, S. S., Wang, P. P., & Te Wang, Y. (2006). A hybrid particle swarm optimization algorithm for optimal task assignment in distributed systems. *Computer Standards & Interfaces*, 28(4), 441–450. doi:10.1016/j.csi.2005.03.005

Ying, L., Shakkottai, S., & Reddy, A. (2009). On combining shortest-path and back-pressure routing over multihop wireless networks. *Proceedings - IEEE INFOCOM*, 1674–1682. doi:10.1109/INFCOM.2009.5062086

Zhang, L., Chen, Y., Sun, R., & Yang, B. (2008). A Task Scheduling Algorithm Based on PSO for Grid Computing. *International Journal of Computational Intelligence Research*, 4(1). Advance online publication. doi:10.5019/j.ijcir.2008.123

KEY TERMS AND DEFINITIONS

Communication Network: Prototype of paths in which information flows in the organization.

Communication Time: In which all users can exchange information instantly or with negligible latency.

Distributed Computing System: Set of processors located at different sites and connected by a communication network.

Execution Time: Time spent by the system executing that task, including the executing running time or system service on its behalf.

Genetic Algorithm: An algorithm that mimics the genetic concepts of natural selection, combination, selection, and inheritance.

Honeybee Mating Optimization: Search algorithm based on swarm approach to optimization.

Load Balancing: Distribution of workloads across multiple computing resources.

Neural Network: Set of algorithms, modelled loosely after the human brain, that is designed to recognize patterns.

Particle Swarm Optimization: Computational method that optimizes a problem by iteratively trying to improve a solution with regard to a given measure of quality.

Response Time: Total amount of time it takes to respond to a request for service.

System Reliability: Describes the ability of a system to function under stated conditions for a given period of time.

Tasks Scheduling: Method by which work is assigned to resources that complete the work.

High-Speed Viterbi Decoder

2

Mário Pereira Véstias

 <https://orcid.org/0000-0001-8556-4507>

INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal

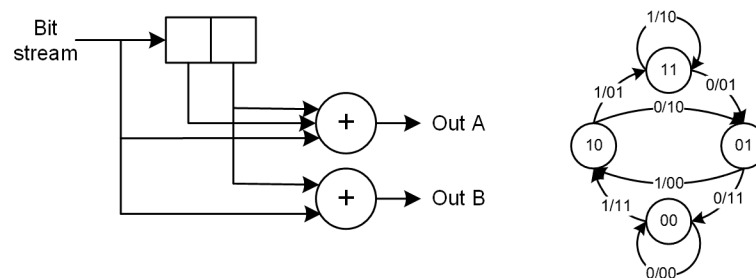
BACKGROUND

Generically, a communication system includes a channel encoder at the transmitter and a channel decoder at the receiver. Encoding the bit stream to be transmitted reduces the energy per data bit needed to achieve the same bit error rate (BER) over a noisy channel compared to a non-coded transmission. A better channel code needs lower energy to obtain a particular BER.

Convolutional Encoder

Two major classes of binary codes are block codes and convolution codes. In this article, the authors are only concerned with convolution codes (P. Elias, 1955). Convolution codes operate over a continuous bit stream. A convolutional encoder is used to generate the encoded bitstream. A typical implementation consists on a shift register and a set of modulo-2 addition. The content of the shift register defines the state of the encoder (s bits), from which a finite state machine with $m=2^s$ states is obtained. Figure 1 shows an example of a convolution encoder with rate 1/2 (rate 1/ c generates c bits for each input bit) and its state machine.

Figure 1. Convolution encoder with rate 1/2



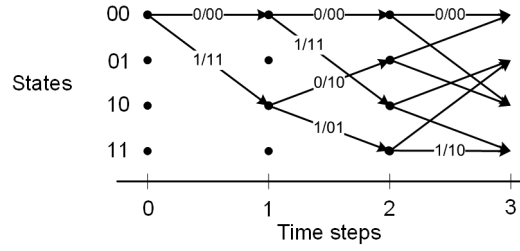
An alternative way to represent the behavior of the encoder was proposed by Forney (G. D. Forney, Jr., 1967) designated *trellis representation*. A trellis is one of the most convenient ways to visualize the behavior of the decoding algorithms (see Figure 2).

Each node in the trellis represents a state of the FSM. After m time steps, the trellis is full and the branch pattern repeats indefinitely.

Most work on trellis decoding was dedicated to the codes with a rate of $1/r$. This type of codes will provide the simplest trellis decoding since there are only two nodes leaving and entering a state.

DOI: 10.4018/978-1-7998-3479-3.ch019

Figure 2. Trellis diagram of the encoder in figure 1



Viterbi Algorithm

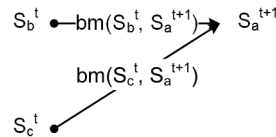
The Viterbi algorithm (VA) was proposed in 1967 by Andrew Viterbi (Viterbi, 1967). The algorithm decodes convolutional codes with an optimum non-sequential algorithm. Contrary to previous algorithmic solutions whose complexity increases exponentially with the length of the encoded sequence, the computational complexity of the Viterbi algorithm increases linearly with the length of the bit stream. The algorithm has three main steps: (1) branch metric calculation; (2) trellis calculation; and (3) trace-back decoding.

The branch metric, bm , calculation determines the squared Euclidian distance, between the received symbol, r_k , and the expected symbol associated with a branch of the trellis diagram, c_k , that is, $|r_k - c_k|^2$. At every time step, $2 \times m$ branch metrics are calculated since each state has two input branches.

The linear computation complexity of the VA is achieved by discarding unlikely branches in the trellis diagram using the Principle of Optimality (Bellman and Dreyfus, 1962). According to this, whenever two paths of the trellis diagram merge into the same state only the path with the best metric (most likely) must be kept. This path is designated *survivor path*. At each time step, this operation is applied to all paths and m survivor paths are obtained at each time step. Formally, a state metric, $sm(state)$ (see figure 3), is updated as follows:

$$sm(S_a^{t+1}) = \min(S_b^t + bm(S_b^t, S_a^{t+1}), S_c^t + bm(S_c^t, S_a^{t+1})) \quad (1)$$

Figure 3. State metric update

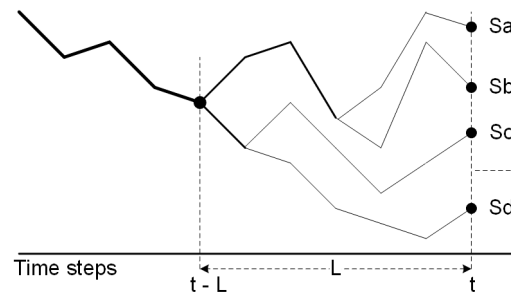


At each time step, the trellis calculation step obtains m decision symbols, one for each surviving branch. These symbols must be stored to help recover the symbols responsible for specific transitions during the traceback decoding step.

The trellis computation stops after a predetermined number of steps or after the decoding of the complete stream. From the m survivor paths, the final step (traceback decoding) starts in the state with the best metric and tracebacks through the path identified by the decision bits until reaching the first state and, consequently, obtain the most likely sequence of symbols.

The higher the number of steps before starting the traceback process, the higher the latency and the memory requirements. To reduce the latency and/or memory requirements the trellis process can be stopped after a pre-defined number of steps by, for example, considering extra bits that terminate the trellis in a defined state. The disadvantage of the method is the overhead associated with the introduction of bits with no information. Fortunately, the VA has one property that allows us to stop the trellis process in the middle of the decoding process and still recover the trellis path. Considering the m survivor paths at a step in the trellis diagram, when traced back, these paths merge into a single path (see Figure 4).

Figure 4. Traceback of survivor paths merging into a single path

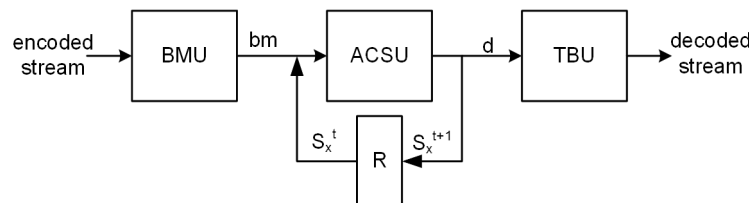


Formal estimations of L can be found in the literature (Anderson and Balachandran, 1989). However, in practice, the required value for L is found by simulation.

Viterbi Decoder

The Viterbi decoder implements the Viterbi algorithm with three functional units, one for each of the three steps of the VA (see Figure 5).

Figure 5. Viterbi decoder



The branch metric unit (BMU) calculates the Euclidian distances associated with each branch for each input symbol. The Viterbi algorithm considers, by definition, that the received sequence is “hard”, that is, the decoded bits are either 0 or 1. Instead of hard decision, soft decision can be considered, where soft bits (several bits) represent real numbers. In practice, at least three bits are used to represent the expected and received symbols.

The branch metrics are then accumulated with the state metrics and the survivor paths are chosen. This is implemented with an add-compare-select unit (ACSU), which adds the metrics, compares the

metrics entering a state and selects the best metric. The decisions, d , are stored in a memory and used by the traceback unit (TBU) to recover the symbols associated with the survivor path.

Considering the branch metric unit implementation, the Euclidian distance between a received symbol, r , and the expected symbol associated with a branch of the trellis diagram, c , is $|r-c|^2$ applied to all dimensions, d . Formally,

$$b_m(r - c) = \sum_{n=0}^{d-1} (r_i^2 - 2r_i c_i + c_i^2)$$

In the equation, both squares, r_i^2 and c_i^2 contribute equally to the equation and thus can be neglected. So, the equation simplifies to:

$$b_m(r - c) = -\sum_{n=0}^{d-1} (r_i c_i)$$

The Add-Compare-Select Unit (ACSU) accumulates the branch metrics of each branch along the trellis diagram. At each time step, it must do m (number of states) accumulations. The ACS unit updates the metric, m_x^t , as follows

$$m_x^t = \min \left(\left(m_{y \rightarrow x}^{t-1} + b_{m_{y \rightarrow x}} \right), \left(m_{z \rightarrow x}^{t-1} + b_{m_{z \rightarrow x}} \right) \right)$$

The ACS unit of a state receives the accumulated metrics of the previous states and the associated branch metrics. After the add-compare-select operations the circuit generates an updated accumulated metric of the survivor path and a decision bit.

There are several implementation alternatives for the ACS unit: (1) direct implementation; (2) serial implementation; and (3) parallel implementation.

In the direct implementation, m add-compare select units are used, one for each of the m states. The direct implementation is a straightforward solution for the VD. The area of the design increases with $m=2^s$. Therefore, a new bit in the encoder state corresponds to an area duplication of a direct ACSU implementation. Besides the exponential increase in the arithmetic units of the ACSU there is also the increase in the number of feedback connections of the ACS units. Usually, the performance of the ACS unit is determined by these feedback connections. Many direct implementations of the VD were proposed for different number of states. For example, in (Lin, et al, 2005) a 64-state direct implementation was proposed and in (Chang, Suzuki, and Parhi, 2000) the proposal was a 256-state direct implementation. The solution is expensive in terms of resources but one time-step takes only the time of a single unit to execute.

A serial implementation of the ACS unit uses less than m ACS units. The execution time of a single time-step depends on the number of available ACS units. In case, for example, a single unit is used, than a time-step takes m times the delay of a single ACS unit. The implementation reduces the utilized resources but reduces the performance compared to the direct implementation. The serial implementation, also designated time-multiplexed solution or folding, reduces the number of ACS units of a direct implementation by assigning multiple add-compare-select operations to a single ACS unit (*folding factor*) (Shung, et al., 1993 – part I), (Shung, et al., 1993, – part II). The time-multiplexed implementation

reduces the number of ACS units, but there is an increased design complexity relative to the controller and an increase in the number of registers. Also, the critical path increases due to extra multiplexers to route data to a single unit which reduces the expected linear increase operating frequency considering the folding factor.

When the encoded data is received in a continuous stream of bits, the path metric which is the accumulated value of the branch metrics, may overflow. To avoid overflow, the path metrics must be normalized. A very efficient technique that avoids overflow is the modulo normalization technique (Shung et al., 1992) that automatically normalizes the path metric. In modular arithmetic the metric, m_j , is replaced by the normalized metric

$$\widehat{m}_j = \left(m_j + \frac{C}{2} \right) (\text{mod } C) - \frac{C}{2},$$

where C is the maximum value represented by the number of bits used for defining the path metric. The differences in the survivor metrics remain unchanged using modular arithmetic provided that $\text{diff} \leq \frac{C}{2}$, where diff is the bound of the difference between survivor metrics. Formally, the required number of bits for the state metric is given by:

$$S_{\text{bits}} = \log_2 (2 \times \Delta_{\text{max}})$$

where $\Delta_{\text{max}} = \lambda_{\text{max}} \log_2 m$, λ_{max} = maximum branch metric and m is the number of states. For example, for 64 states, and code rate of 1/3, $\lambda_{\text{max}} = 18$. So, $S_{\text{bits}} = 8$. This normalization is implemented in hardware using a 2's complement adder.

Another simplification rule following the metric calculation using modular arithmetic is the modified compare (Shung et al., 1992). The modified comparison rule applied to two normalized metrics with p bits

$$\widehat{mx} = \left(\overline{mx_{p-1} + mx_{p-2} + \dots + mx_0} \right) \text{ and } \widehat{my} = \left(\overline{my_{p-1} + my_{p-2} + \dots + my_0} \right)$$

is given by:

$$mx_{p-1} \oplus my_{p-1} \oplus y(mx_{p-2} + \dots + mx_0, my_{p-2} + \dots + my_0)$$

where $y(\cdot)$ denotes the unsigned comparison. In general, this comparator is smaller and faster than using a subtractor to implement comparison.

The trace-back unit receives the accumulated metric values of all states, finds the state with the lowest value and returns the path ending in this state, that is, the sequence of bits that best matches the received sequence. A fast approach is to use a tree of comparators having $\log_2 m$ levels, that is, $\frac{m}{2}$ comparators at the first level, $\frac{m}{4}$ at the second, and so on (or $\frac{m}{2}$ shared comparators). After $\log_2 m$ cycles, the state with the lowest metric is identified. The solution is fast but expensive in terms of hardware

resources. An approach with small area is to consider a recursive process where at each iteration the highest value found so far is compared to a next value. This sequential solution takes m cycles to complete, increasing considerably the latency and may become the bottleneck if the ACS step takes less than m cycles. Usually, an efficient solution is to consider a tree of iterative comparators (Véstias et al., 2012).

TECHNIQUES FOR HIGH-SPEED VITERBI DECODING

Several techniques to speed-up Viterbi decoding exist at different levels: bit-level, word-level and algorithmic level.

Bit-Level Optimizations

The lowest level optimizations introduces parallelism at the bit level. These include improved techniques for addition (e.g., carry-save adders) and hardware optimizations and reorganization inside BM and ACS units.

Considering noiseless symbols 0 and 1 of c (see previous section) represented by the symmetrical values -1 and +1, the branch metric unit reduces to add/subtraction of the received values

$$b_m(r - c) = -\sum_{n=0}^{d-1} \pm r_i$$

Therefore, only additions need to be implemented to calculate the branch metrics. The number of operations can be further reduced to half since $b_m = -\overline{b_m}$.

A few bit-level optimizations techniques have also been proposed to improve the performance of the ACS unit. For example, an ACS parallelization technique was proposed in (Véstias & Sarmiento, 2010) to reduce the critical path of the unit. The critical path of the ACS unit includes one adder, one comparator and one multiplexer. To reduce this critical path, the addition in the ACS unit is done in parallel with the comparison.

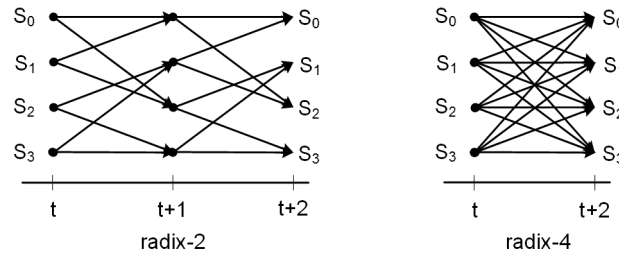
Word-Level Optimizations

Word-level optimizations merge consecutive transitions of the trellis diagram into a single transition to reduce the number of iterations. To determine the minimal path, transitions are unmerged (Cheng, C. & Parhi, K., 2008).

A lookahead technique of two or more levels is a quite effective parallelization technique (Fettweis, 1989), (Black & Meng, 2002). In a single cycle the technique updates the state metrics of n consecutive steps and thus ideally increases the throughput n times, compared to a direct implementation. However, there is an area penalty proportional to n . To see how the technique works, let's consider a trellis diagram with states with two inputs and two outputs (radix-2) and collapse two steps into a single one producing a radix-4 trellis diagram (four branches entering and leaving each state) (see Figure 6).

Higher radices could be considered. However, as concluded in (Black & Meng, 1992), radix-4 solutions are more area efficient than higher radices. Also, the critical path increases due to the extra operations of the ACS unit and the feedback interconnections. In (Black & Meng, 1992), a full-custom

Figure 6. A radix-2 trellis diagram transformed into a radix-4 trellis



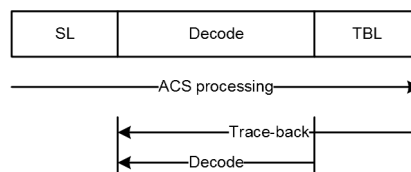
implementation of both radix-2 and radix-4 architectures have shown speed-ups of 1.7, lower than the theoretical 2 times speed-up.

Algorithmic-Level Optimizations

Unbounded throughput using parallelization can be obtained with the minimized method (Fettweis & Meyr, 1991) and the sliding-block decoder (Black & Meng, 1997). Following these techniques, parallel implementations of the VD can be implemented by decomposing the input data into $k \times M$ blocks which can be processed in parallel using k conventional Viterbi decoders. The complexity of the circuit increases linearly with the throughput. However, each block processing requires the initial state metrics to be known and this is only available after the previous block has finished processing. Three techniques have been proposed as a solution to this problem: state initialization (Dunn, 1978), interleaving (Lin & Messerschmitt, 1989) and sliding block (Tzu & Dunham, 1981). The state initialization technique forces the encoder to a known state at the start of each block. This adds complexity to the decoder since it needs block synchronization. The interleaving technique processes independently encoded sequences. However, this approach cannot be used when intersymbol interference is present. The sliding block decoder technique processes independent overlapping blocks simultaneously. To acquire the initial states, two sync blocks are used at the beginning and at the end of the block. The method is based on two main observations. One is that the survivor paths from all possible starting states merge with high probability L iterations back into the trellis. The parameter is the survivor path length (SPL) and is typically $5 \times s = 5 \times \log_2 m$ (Clark & Cain, 1981). The other is that when starting with unknown initial state metrics, the state metrics after K trellis iterations are independent of the initial metrics. The parameter is the synchronization length (SL) and is also typically $5 \times \log_2 m$ (Viterbi & Omura, 1979).

Based on this characteristics, a state at time n can be decoded using only information from the interval $n-K$ to $n-L$. Typically, $K = L$ is considered. The sliding block VD (SBVD) method applies the Viterbi algorithm to the interval $[n - M/2 - L, n + M/2 + L]$ to decode the interval $[n - M/2, n + M/2]$. To guarantee correct initial and final states, the method considers an initial synchronization block (SL)

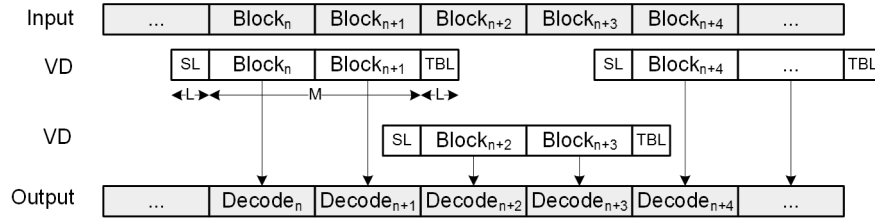
Figure 7. Sliding block method for Viterbi decoder



and a final trace-back block (SPL) without decoding for each parallel VD (see the behavior of each VD for the SBVD in Figure 7).

With this technique, independent parallel VD can be used to increase the decoding rate (see Figure 8).

Figure 8. Sliding block method for Viterbi decoder



The relative area efficiency is $\frac{M}{M + 2L}$. As M increases the method approaches the ideal efficiency of 1. However, the bigger the value of M, the bigger the latency and more storing resources are needed to save intermediate states. The decoding rate, DR, of an SBVD is:

$$DR = nVD \times \frac{M}{M + 2L} \times drVD$$

where nVD is the number of parallel VD and $drVD$ is the decoding rate of a single VD. The bit error rate of the decoding process of the VD depends on the length of the initial synchronization step, the length of the final trace-back length and of the trace-back decode length. All these parameters are usually found through system simulation.

In (Sun, F. & Zhang, T., 2007) the optimization process considers the elimination of transitions that are bigger than the current minimum reducing the number of operations to be executed.

FUTURE RESEARCH DIRECTIONS

Fast Viterbi decoders are fundamental for high speed applications that may be subject to real-time constraints. Parallel techniques allow any decoding rate just by increasing the number of parallel Viterbi decoders and thus an increase in the area and power consumption.

One of the characteristics of the Viterbi decoder is that of being the most resource-consuming compared to other decoding algorithms, but theoretically it does maximum likelihood decoding. However, there are some obstacles to yield the ideal maximum likelihood due to several factors, including quantization of the input stream and finite trace back length.

A major challenge in the future of the Viterbi decoder is how to achieve high throughput with low power consumption. This is very important for portable devices.

CONCLUSION

The article describes the backgrounds of the Viterbi algorithm as a way to design a Viterbi decoder. The typical architecture of the Viterbi decoder is explained. Several techniques to improve the decoding speed of Viterbi decoders were described and discussed, including bit, word and algorithmic level techniques.

The Viterbi algorithm was proposed in 1967 and since then many Viterbi decoder architectures were proposed. Different design requirements are appearing every day in different technologies and different applications, in particular, mobile devices that require high-speed at low cost and low energy consumption. Designing high-speed Viterbi decoders is therefore an actual topic of research.

REFERENCES

- Anderson, J. B., & Balachandran, K. (1989). Decision depths of convolutional codes. *IEEE Transactions on Information Theory*, 35(2), 455–459. doi:10.1109/18.32141
- Bellman, R. E., & Dreyfus, S. E. (1962). *Applied Dynamic Programming*. Princeton University Press. doi:10.1515/9781400874651
- Black, P., & Meng, T. (2002). 140-Mb/s, 32 state, radix-4 Viterbi decoder. *IEEE Journal of Solid-State Circuits*, 27(12), 1877–1885. doi:10.1109/4.173118
- Black, P. J., & Meng, T. H.-Y. (1992). A 140-Mb/s, 32-state, radix-4 Viterbi decoder. *IEEE Journal of Solid-State Circuits*, 27(12), 1877–1885. doi:10.1109/4.173118
- Black, P. J., & Meng, T. H.-Y. (1997). A 1-Gb/s, four-state, sliding block Viterbi decoder. *IEEE Journal of Solid-State Circuits*, 32(6), 797–805. doi:10.1109/4.585246
- Chang, Y.-N., Suzuki, H., & Parhi, K. K. (2000). A 2-Mb/s 256-state 10-mW rate-1/3 Viterbi decoder. *IEEE Journal of Solid-State Circuits*, 35(6), 826–834. doi:10.1109/4.845186
- Cheng, C., & Parhi, K. (2008). Hardware efficient low latency architecture for high throughput rate viterbi decoders. *IEEE Transactions on Circuits and Systems, II, Express Briefs*, 55(12), 1254–1258. doi:10.1109/TCSII.2008.2008061
- Clark, G. C., & Cain, J. B. (1981). *Error-Correction Coding for Digital Communications*. Plenum. doi:10.1007/978-1-4899-2174-1
- Dunn, J. (1978). A 50 Mb/s Multiplexed Coding System for Shuttle Communications. *IEEE Transactions on Communications*, COM-26, 1877–1638.
- Elias, P. (1955). Coding for noisy channels. *IRE Convention Record*, 2, 37–47.
- Fettweis, G., & Meyr, H. (1989). Parallel Viterbi algorithm implementation: Breaking the ACS bottleneck. *IEEE Transactions on Communications*, 37(8), 785–790. doi:10.1109/26.31176
- Fettweis, G., & Meyr, H. (1991). High-speed parallel Viterbi decoding: Algorithm and VLSI-architecture. *IEEE Communications Magazine*, 29(5), 46–55. doi:10.1109/35.79382
- Forney, G. D., Jr. (1967). *Review of random tree codes*. Final Report Appendix A Contract NAS2-3737, NASA CR 73176. NASA Ames Research Center.

- Lin, C.-C., Shih, Y.-H., Chang, H.-C., & Lee, C.-Y. (2005). Design of a power reduction Viterbi decoder for WLAN applications. *IEEE Transactions on Circuits and Systems—Part I: Regular Papers*, 52(6), 1148–1156. doi:10.1109/TCSI.2005.849106
- Lin, H., & Messerschmitt, D. (1989). Algorithms and Architectures for Concurrent Viterbi Decoding. *Proceedings of International Conference on Communications*, 2, 836–840. 10.1109/ICC.1989.49807
- Shung, C. B., Lin, H.-D., Cypher, R., Siegel, P. H., & Thapar, H. K. (1993). Area efficient architectures for the Viterbi algorithm—Part I: Theory. *IEEE Transactions on Communications*, 41(4), 636–644. doi:10.1109/26.223789
- Shung, C. B., Lin, H.-D., Cypher, R., Siegel, P. H., & Thapar, H. K. (1993). Area efficient architectures for the Viterbi algorithm—Part II: Applications. *IEEE Transactions on Communications*, 41(5), 802–807. doi:10.1109/26.225495
- Shung, C. B., Siegel, P. H., Ungerboeck, G., Thapar, H. K., Trung, T. K., Shih, M. T., Reed, I. S., & Satorius, E. H. (1992, March). VLSI architectures for Metric Normalization in the Viterbi algorithm design. *IEEE Transactions on Communications*, 40(3), 616–624.
- Sun, F., & Zhang, T. (2007). Low-power state-parallel relaxed adaptive viterbi decoder. *IEEE Transactions on Circuits and Systems. I, Regular Papers*, 54(5), 1060–1068. doi:10.1109/TCSI.2007.890617
- Tzu, K.-H., & Dunham, J. (1981). Sliding Block Decoding of Convolutional Codes. *IEEE Transactions on Communications*, COM-29(September), 1401–1403.
- Véstias, M., Neto, H., & Sarmiento, H. (2012). Efficient Design of Sliding Block Viterbi Decoders in FPGA. *Proceedings of IEEE International Conference on Field Programmable Logic Applications*.
- Véstias, M., & Sarmiento, H. (2010). Implementing High-Speed Viterbi Decoders for MB-OFDM on FPGA. *Proceedings of XXV Conference on Design of Circuits and Integrated Systems*.
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260–269. doi:10.1109/TIT.1967.1054010
- Viterbi, A. J., & Omura, J. K. (1979). *Principles of Digital Communication and Coding*. McGraw-Hill.

ADDITIONAL READING

- Anders, M. A., Mathew, S. K., Hsu, S. K., Krishnamurthy, K., & Borkar, S. (2008). A 1.9 Gb/s 358 mW 16–256 State Reconfigurable Viterbi Accelerator in 90 nm CMOS. *IEEE Journal of Solid-State Circuits*, 43(1), 214–222. doi:10.1109/JSSC.2007.909336
- Anderson, J. B. (2005). *Digital Transmission Engineering* (2nd ed.). IEEE Press. doi:10.1002/0471733660
- Asher, Y., Tartakovsky, V., Portman, K., Zilberman, O., & Hadar, A. (2018). An FPGA Scalable Parallel Viterbi Decoder. *IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoc)*, 8–15.
- Backurs, A., & Tzamos, C. (2016). *Improving Viterbi is Hard: Better Runtimes Imply Faster Clique Algorithms*. <https://arxiv.org/abs/1607.04229>

Balivada, L. K., & Raju, K. P. (2012). *Optimization Techniques of Viterbi Algorithm: Performance Analysis of Different Algorithms*. LAP LAMBERT Academic Publishing.

Benedetto, S., & Biglieri, E. (2002). *Principles of Digital Transmission: With Wireless Applications*. Springer US.

Chu, C.-Y., & Wu, A.-Y. (2012). Power-Efficient State Exchange Scheme for Low-Latency SMU Design of Viterbi Decoder. *Journal of Signal Processing Systems for Signal, Image, and Video Technology*, 68(2), 1–13. doi:10.1007/11265-011-0603-0

Kamuf, M., Owall, V., & Anderson, J. (2008). Optimization and Implementation of a Viterbi Decoder Under Flexibility Constraints. *IEEE Transactions on Circuits and Systems I*, 55(8), 2411–2422. doi:10.1109/TCSI.2008.918148

Kermani, M., Singh, V., & Azarderakhsh, R. (2017, January). Reliable Low-Latency Viterbi Algorithm Architectures Benchmarked on ASIC and FPGA. *IEEE Transactions on Circuits and Systems. I, Regular Papers*, 64(1), 208–216. doi:10.1109/TCSI.2016.2610187

Lin, M.-B. (2000). New Path History Management Circuits for Viterbi Decoders. *IEEE Transactions on Communications*, 48(10), 1605–1608. doi:10.1109/26.871381

Lin, S., Kasami, T., Fujiwara, T. & Fossorier, M. (1998). Trellis and Trellis-Based Decoding Algorithms for Linear Block Codes. *Springer Science+Business Media, LLC*.

Mandwale, A., & Mulani, A. (2015). Different Approaches For Implementation of Viterbi decoder on reconfigurable platform. *International Conference on Pervasive Computing (ICPC)*, pp. 1-4. 10.1109/PERVASIVE.2015.7086976

Middya, A., & Dhar, A. S. (2016). Real-time area efficient and high speed architecture design of Viterbi decoder. *2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 246-250. 10.1109/AEEICB.2016.7538282

Mostafa, K., Hussein, A., Youness, H., & Moness, M. (2016). High performance reconfigurable Viterbi Decoder design for multi-standard receiver. *33rd National Radio Science Conference (NRSC)*, pp. 249-256. 10.1109/NRSC.2016.7450835

Peng, H., Liu, R., Hou, Y., & Zhao, L. (2016). A Gb/s parallel block-based Viterbi decoder for convolutional codes on GPU, *8th International Conference on Wireless Communications & Signal Processing (WCSP)*, pp. 1-6. 10.1109/WCSP.2016.7752638

Pradhan, A., & Nandy, S. (2014). A reconfigurable viterbi decoder for SDR and mobile communications. *International Conference on High Performance Computing and Applications (ICHPCA)*, 2014, pp. 1-6. 10.1109/ICHPCA.2014.7045373

Prakash, S., & Balamurugan, V. (2016). Design and implementation of Fast Radix-2 and radix-4 ACSU with different adders in Viterbi decoder. *International Conference on Green Engineering and Technologies (IC-GET)*, pp. 1-4.

Sugur, N., Siddamal, S., & Vemala, S. (2014). Design and Implementation of High Throughput and Area Efficient Hard Decision Viterbi Decoder in 65nm Technology,” *27th International Conference on VLSI Design and 13th International Conference on Embedded Systems*, pp. 353-358.

Swathi, I., & Rajaram, S. (2017). FPGA implementation of viterbi decoder for software defined radio applications. *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2070-2073. 10.1109/WiSPNET.2017.8300126

KEY TERMS AND DEFINITIONS

Add-Compare-Select Unit (ACSU): A unit that accumulates the branch metrics of each branch along the trellis diagram.

Branch Metric Unit (BMU): A unit to calculate the Euclidian distances associated with each branch of the Trellis diagram for each input symbol.

Convolutional Codes: A type of error correcting code used in telecommunications.

Convolutional Encoder: A module used to generate convolutional codes.

Trace-Back Unit (TBU): A unit of the VD that finds the state with the lowest value and returns the path ending in this state.

Trellis Diagram: A type of graph where nodes represent states and each node is connected to at least one previous node (state) and one next node (state).

Viterbi Algorithm: An algorithm to decode convolutional codes with an optimum non-sequential algorithm. The computational complexity of the Viterbi algorithm increases linearly with the length of the bit stream. The algorithm has three main steps: branch metric calculation; trellis calculation; and traceback decoding.

Viterbi Decoder (VD): An implementation of the Viterbi algorithm consisting of three main functional units, one for each of the three steps of the Viterbi algorithm.

Field-Programmable Gate Array

2

Mário Pereira Véstias

 <https://orcid.org/0000-0001-8556-4507>

INESC-ID, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal

INTRODUCTION

Field-Programmable Gate Arrays (FPGAs) are integrated circuits whose logic and their interconnections are configurable. These devices are field-programmable, that is, they can be configured by the hardware designer without any intervention of the manufacturer. Most FPGAs can be reprogrammed as many times as we want with a vast variety of digital circuits.

The applicability of FPGAs have changed since they appeared in the mid-1980s. Initially they were very limited in terms of resources and therefore were used to implement glue logic and small sized digital systems. With the advance of the integrated circuit technology, its density has increased dramatically permitting the implementation of a vast set of applications in different areas like consumer, industrial, automotive, high-performance and communications. Some recent FPGA families are system-on-chips (SoC) with one or more microprocessor cores, memory, cache and reconfigurable logic allowing the implementation of complex hardware/software systems in a single programmable device.

FPGAs are now an alternative for ASIC (Application Specific Integrated Circuits). ASIC are the best solutions for implementing digital systems since they offer better performance and lower power consumption at smaller silicon area. The problem is that designing an ASIC is very time-consuming, with high nonrecurring engineering (NRE) costs requiring expensive tools to design and develop. Besides, after fabrication of the final design in ASIC it cannot be modified without creating a new chip.

The cost unit of an ASIC is lower than that of an FPGA but has a much higher NRE cost. Therefore, the ASIC is the best option only for high volume products. This makes FPGAs the chosen device for small to medium scale product deployment. Designing an FPGA is easier, it permits modifications and upgrades of the design, has a faster time-to-market and so the only viable solution for small companies that cannot incur the high NRE costs and expensive design tools.

This article focus on the architecture of FPGAs, including the so called SoC FPGA. It explains the main blocks of the FPGA, how they have evolved along the last decades and the perspectives of next generation FPGAs. It also describes some applicability areas and how its architecture have evolved to adapt to some of these target markets.

BACKGROUND

Field-Programmable Gate Arrays are silicon devices whose fabric hardware can be programmed to implement a particular digital system. The hardware programming capability of FPGAs gives them some advantages over Application Specific Integrated Circuits (ASIC) whose underlying fixed hardware is design to implement a specific digital system. Deploying and maintaining an FPGA-based system is faster and cheaper than implementing the same system in ASIC, since the ship is already made and just

DOI: 10.4018/978-1-7998-3479-3.ch020

have to be configured to implement a specific function. Later, this function can be changed in the field without having to redesign a new chip and can even be reconfigured dynamically while running some other functionality. While is cheaper to design FPGA than an ASIC, the unit cost is higher for an FPGA. Therefore, the average cost of each of these devices depend on the production volume. The average unit cost of an ASIC decreases faster with the production volume than the cost of an FPGA unit. Therefore, there is a threshold from which ASICs are more cost effective.

The programming flexibility of FPGAs comes with a cost in terms of performance and power consumption. A large area of the FPGA (from 80 to 90%) is dedicated to routing and programming resources. It means that only 10-20% of the chip area is available to implement the hardware system. So, compared to an ASIC, a larger FPGA is necessary to implement the same functionality, the performance is worst due to the delay associated with the configurable routing and logic and the higher power consumption includes the power consumption of programming resources.

Any digital system consists of functional blocks, interconnections between these blocks and input/outputs connections between the functional blocks and off-chip circuits. An FPGA consists of programmable logic, interconnect and I/O modules that are tailored to implement a specific digital system.

FPGA Functional Blocks

The granularity of functional blocks determines the flexibility of the FPGA. The smaller the granularity the more flexible it becomes and the FPGA can be more deeply adapted to the required logic system. In the limit, we can use transistors as the basic cells that could then be interconnected to implemented basic logic gates. This approach was followed in (Marple, D., & Cooke, L., 1992) without success mainly because this very fine granularity requires large amounts of interconnections turning the device quite area-inefficient, with large propagation delays and high power consumption associated with the capacitances of the programmable interconnections.

Different granularities were tried in the last decades, from NAND gates (Plessey, 1992), trees of multiplexers (Gamal et al., 1989) and lookup tables (Carter et al., 1986). Today, an FPGA consists of fine-grain configurable circuits of logic, I/O and interconnections arranged in a regular map see Figure 1).

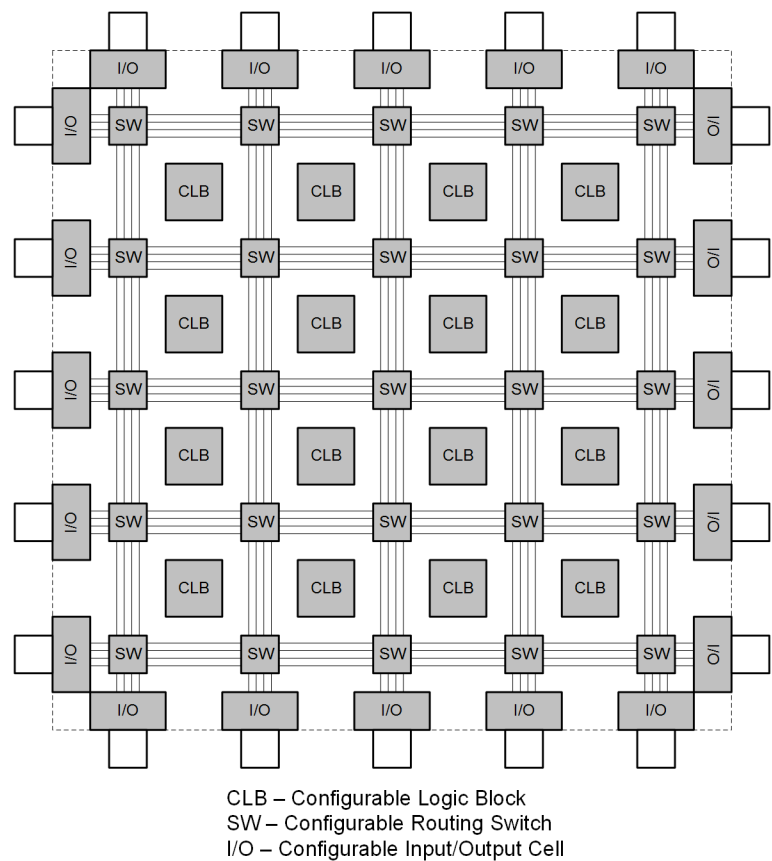
Configurable input/output (I/O) blocks establish the communication of data to and from the system configured in the FPGA. Configurable logic blocks (CLB) implement the functionality of the system and routing switch (SW) provide configurability to the routing resources.

A configurable logic block issues fine-grained configurable logic and memory resources organized in a hierarchical structure (see figure 2).

A configurable logic block (different vendors call different names to these blocks: CLB, LAB – logic array block, etc.) encapsulates fine-grained logic and memory and some specialized configurable circuitry that improves the design of some functions, like arithmetic. One CLB contains one or more cells with dedicated connections between neighbor cells (the designation of these elements are also vendor dependent; e.g. ALM – adaptive logic module and slice).

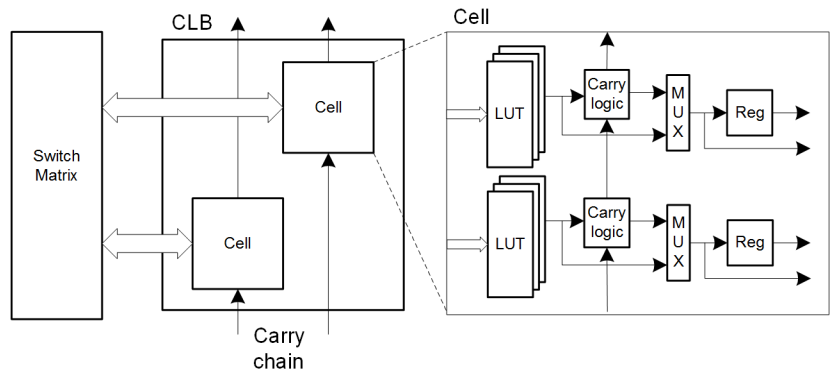
Logic in FPGAs is implemented with small look-up tables (LUT) which are basically small memories whose content can be changed to implement a specific function. A LUT with k inputs, designated LUT- k , implements any boolean function with k -inputs. Each cell of a CLB contains one or more locally interconnected LUTs and flip-flops. The size of the LUT determines the area and the delay of the circuit according to two trading dimensions. Larger LUTs reduce the number of logic blocks in the critical path, the amount of inter logic routing, which improves performance. However, as we increase the size of the LUT, the internal delay increases and so smaller logic functions mapped to the LUTs have higher delays.

Figure 1. Basic underlying architecture of an FPGA



Since LUTs are implemented as small memories, vendors took advantage of this to increase the configurability and functionality of slices. LUTs can be used not only to implement logic, but also as small distributed memories or in some cases as shift registers. This flexibility permits the implementation of memories and shift registers without using the flip-flops present in a slice which are left for other functionality.

Figure 2. Hierarchical structure of a fine-grained CLB



FPGA Routing

CLB cells are connected to each other to implement larger and more complex logic functions. Interconnections between CLBs and I/O blocks are implemented with a programmable routing network consisting of wires and switches to route data through these wires. To be flexible enough and accommodate large and complex circuits, routing interconnections must be massive and flexible.

The propagation delay of a wire increases with its length. Fortunately, most designs exhibit locality and so fast local and short wires should be more abundant than long wires for a more efficient interconnection network. The network must also consider the special clock and reset lines and distribute them across the FPGA so that clock signals are distributed with low skew and able to keep sequential circuits synchronized.

The architecture of the interconnection network defines the position of the channels in between the sea of logic blocks, the width of channels and how they interconnect to each other. At a more detailed level, it defines the length of wires, its locality, the switching patterns and how they connect to the logic blocks. The global organization of routing architectures can be of two different types: hierarchical (Aggarwal, A. & Lewis, D., 1994) or island-style (Brown, S. D., 1992). Hierarchical routing organizes blocks of the FPGA into clusters. Logic blocks inside a cluster are interconnected with local wires, while interconnecting blocks in different clusters must go through one or more levels of routing. The structure is similar to a tree of connections where the higher the level the more wires are in a channel. The problem with this type of organization is in the design mapping since an unmatched between the routing architecture and the routing requirements of the design forces the underutilization of logic blocks in a cluster. Also, routing between blocks in different clusters incur a large delay penalty. These disadvantages took commercial FPGAs to adopt a different routing architecture where there exists only one flat level of routing channels – island-style routing.

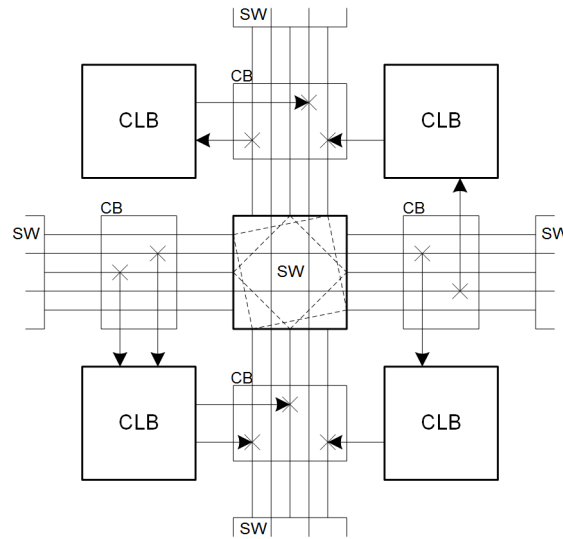
The island-style routing organizes the logic blocks in a two dimensional array surrounded by routing channels (see figure 1 above). A channel has segments of different lengths which are used according to the connection requirements of each logic block. The locality of communications between logic blocks determined that short, local wires are more abundant than long wires in a channel. This routing style has more positive features than hierarchical routing. Logic blocks have direct access to wires of different lengths improving the connection efficiency of designs requiring different routing patterns. Also, the layout of the island-style structure is more regular improving the scalability of the architecture when compared to a design with a hierarchical routing.

The island-style routing architecture also determines how logic blocks connect to the interconnection structure and how channels interconnect to each other (see figure 3).

Each output and input pin of a logic block connects to the channels through input and output channel blocks. Pins of a logic block do not connect to all wires of a channel block (CB). Connections between adjacent channels are made with programmable switches (SW). These modules have configurable connections that implements different connection patterns whose flexibility depends on the adopted design. Extensive studies have analyzed and looked for the best wire segments size (Betz, V. & Rose, J., 1999), the best switch architectures (Masud, M. I. & Wilton, S., 1999) and the best block channels.

The layout of wires and the type of switch also determines the performance of the interconnection network. Pass transistors, multiplexers and buffers are used to implement basic switches where different ratios between the numbers of each of these components determines the fastest routing structure (Betz, V. & Rose, J., 1999).

Figure 3. Low-level structure of the island-style routing network



Input/Output

Input/output cells of FPGAs interface with external components. Many different external devices must interact with circuits inside the FPGA and so input/output cells are also configurable to guarantee enough interface flexibility that broadens the range of devices that can be connected to the FPGA.

The challenge in designing input/output cells is that many I/O standards exist with different voltage levels (e.g., 1.2V, 1.8V, 3.3 V), transmission speeds, signaling, etc. Supporting as much standards as possible gives FPGAs a great interface flexibility but also increases the I/O silicon area and increases pin capacitance with consequences on I/O performance. In general, the choice of which standards to support depends on the target markets of the FPGAs. Even with a limited set of standards, implementing all I/O with support for all interfaces in this set is not area and performance efficient. Instead, FPGA vendors adopted heterogeneous I/O cells with different banks of cells supporting different sets of standards. This reduces the complexity and improves performance while still guaranteeing for most cases the diversity of I/O standards.

Heterogeneity in I/O cells organized in banks of I/O open a design space relative to the number and type of banks. There are FPGAs where the number of I/O is almost constant for all banks (Xilinx, 2005), while others consider a fixed number of banks (Altera, 2005). Recent FPGAs have adopted a hybrid approach with different FPGA devices having a different number of banks and a variable number of pins per bank (Altera, 2006) providing a more balanced solution in terms of area, complexity and performance.

The speed of I/O is another important aspect in the design of the FPGA. Access to high-speed memory and high-speed links require additional and specific circuitry to support high speed communications. Some high speed interfaces require differential signaling that is implemented with two I/O paired together. Support for multi-gigabits interfaces requires specialized modules for data serialization.

Programming Technology

Different technologies are used to store the programming bits of reconfigurable devices. The most well-known and used is SRAM (Static Random Access Memory) (Carter et al., 1986), flash (Guterman et al., 1979) and anti-fuse (Birkner et al., 1992). In the SRAM technology, program memory cells are implemented with SRAM cells, that is, to control select inputs of routing and logic multiplexers and to implement LUTs. SRAM-based programming technology is the preferred among many FPGA providers. SRAM cells can be reprogrammed an infinite number of times, are fast and are implemented with CMOS technology the same used to implement the logic circuitry of the FPGA. Therefore, the fabrication process is the same for the whole FPGA. However, the technology has some drawbacks, including the size (five or six transistors), volatility (external non-volatile memories are required to store the programming file for the FPGA), security (the configuration bitstream must be loaded into the FPGA device from an external memory which increases the risk of being copied. Encryption of the bitstream is used to avoid security problems) and capacitive load of pass transistors that increases with the reduction of integrated circuit technology.

Some of these disadvantages can be addressed using the flash technology. Flash is non-volatile which eliminates the necessity to load the configuration from an external memory. A programmed flash-based FPGA is ready to work as soon as it is powered-up since it does not have to load the configuration. It is also smaller than SRAM since only needs one transistor but requires high voltage buffers to program flash cells, which increases the area. Contrary to SRAM technology, flash cells cannot be reprogrammed an infinite number of times and requires a different fabrication process.

Anti-fuses are structures with high resistance that can be programmed to have a low resistance. However, unlike the two previous technologies, an anti-fuse can be programmed only once. One of the main advantages of this technology is that the connections can be programmed without silicon, only using metal and the memory is non-volatile. However, it requires current sources to program the anti-fuses. Anti-fuses occupy low area compared to SRAM and flash memories. The main problem associated with anti-fuses is that the manufacturing requires a non-CMOS technology. Also, the technology is not scalable to the recent technology sizes, while CMOS is always up to date, and programming cannot be done without removing the device from the board.

The Road of SRAM-based Field-Programmable Gate Arrays

Since they were introduced in 1984, FPGAs have gone a long evolution process dictated by IC technology evolution, architectural evolution looking to improve performance, silicon area utilization and power consumption, and architectural modifications determined by the target applications. In the following sections, we describe the roadmap of evolution of FPGAs driven by these aspects.

Technology Roadmap of FPGAs

The first FPGA device was proposed by Wahlstrom in 1967 (Wahlstrom, S. E., 1967). The device consisted of a regular array of logic cells that could be programmed to implement logic functions and memory interconnected by a configurable network. Configuration bits were stored in SRAM memory cells. While flexible, an SRAM cell is expensive in terms of silicon area and most of the FPGA silicon was devoted SRAM to store the programming bits. This silicon overhead has delayed the introduction of a commercial FPGA until 1984 when Xilinx introduced the first FPGA – XC2064 (Carter et al., 1986).

Since then the evolution of integrated circuit (IC) technology has allowed high density programmable devices where full hardware/software systems can now be implemented.

The first FPGA – XC2064 – had 64 logic blocks, each with two 3-input LUTs and one register. The die area pushed the IC technology to the limits increasing the cost of the device. The programmability was based on SRAM which requires an external non-volatile memory to store the configuration stream. To face these disadvantages, antifuse-based FPGAs eliminate the need of an external memory and reduces the size of programming memory, at the cost of only one-time programmability. Actel 1280 was one of the first FPGA with anti-fuse technology in 1990.

This scenario suffered a turn around with the great evolution of semiconductor technology following Moore's Law. In 1992 the XC4010 FPGA from Xilinx had about 1000 gates and in 1999 this figure raised to about 1 million with the XCV1000 FPGA with a 0.18 μm technology. This large density fabrication technology brought SRAM-based technology back not only because now silicon area was not a problem but also because anti-fuse technology was unable to keep with the CMOS process technology. Using a homogeneous IC technology was necessary for FPGAs to progress at the same pace of IC fabrication processes. Besides, reprogrammability is now an important aspect that permits in-system programmability and system upgrade.

The IC technology keeps its evolution and FPGAs are at the forefront to use them. The 14 nm Tri-Gate process from Intel was used to fabricate the Stratix 10 Family of FPGAs from Intel (Intel, 2019). This constant increase in IC density drove the design decisions of the architecture of FPGAs looking for the best performance and area efficiencies.

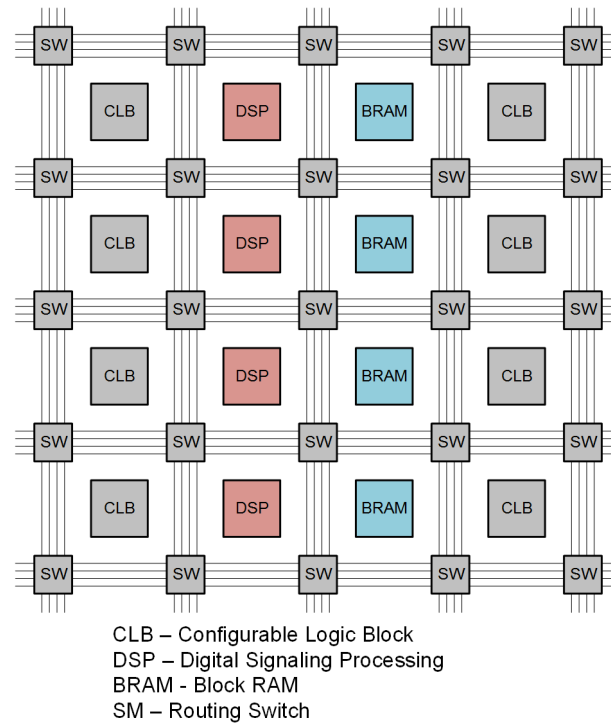
Designing FPGAs for Best Efficiency

The baseline structure of an FPGA considers fine-grained homogeneous configurable logic blocks. Fine-grained logic is quite adaptable to a broad set of logic systems but implies a large fabric overhead in terms of programming circuits for logic and routing and a penalty in performance because of programming switches. On the other side, dedicated logic with specific functionality improves performance, area and power consumption but reduces the mapping opportunities, that is, if a target system does not require this functionality then the silicon area of the dedicated logic is wasted. This creates a tradeoff that has been explored by FPGA vendors.

An important design aspect of FPGAs that determines the area efficiency is the size of a LUT. The tradeoff between the size of a LUT and the area efficiency has been extensively studied (Ahmed & Rose, 2000), (Ahmed & Rose, 2004). The first FPGA used a 3-input LUT. With the increase in IC density LUTs have increased to 4-inputs, LUT-4, and are now with 6-inputs, LUT-6. With more silicon area, keeping the size of the LUTs means more routing silicon area. Larger LUTs reduce the number of logic blocks in the critical path, the amount of inter logic routing, improving performance. However, LUTs with larger granularity have higher internal delays increasing the delay of smaller logic functions. FPGA vendors have introduced some programmability to the size of LUTs. For example, a 6-input LUT of a Xilinx FPGA can be configured as two 5-input LUTs with common inputs or one 6-input LUT. The Altera Stratix FPGA uses ALM with 8-input LUTs that can implement one 6-input logic function, two 4-input logic functions, one 5-input and one 3-input functions, or two input functions that share the same logic function and four inputs.

Another architectural approach to improve area efficiency is to consider dedicated circuitry or hardware modules. Dedicated hard modules with a custom functionality implements this functionality more

Figure 4. Heterogeneous architecture of an FPGA with multiple dedicated circuits



efficiently than using fine-grained configurable logic blocks. The disadvantage is that if the applications does not require them then its silicon area is wasted.

Two levels of dedicated circuits exist in modern FPGAs. The lowest is at the level of the configurable logic block where some circuitry is dedicated to improve arithmetic operations (dedicated circuits for carry chain, bit multiplication, XOR gates for addition) and some for memory (shift registers, support for first-in-first-out buffering, RAM and ROM memories). At a higher level, medium-grained blocks independent of the configurable logic blocks exist also in a tiled regular structure mixed with CLBs (see figure 4).

Block RAM (BRAM) first appeared in a commercial FPGA (Altera, 2003). This memory blocks have a fixed memory capacity but the number of address and data bits is configurable. A set of different aspect ratios are available (e.g., a BRAM with 2K bits can be configured as $2K \times 1$, $1K \times 2$, 512×4 , 256×8 , 128×16 , or 64×32). These blocks can be clustered together to obtain larger memories, in some cases without extra fabric logic. Some FPGAs have heterogeneous memory sizes, with a few extra larger memories (e.g. 512 Kbit) targeting applications where large on-chip memory is important.

Another important configurable aspect of BRAMs in modern FPGAs is the number of ports where single or dual ports are available. Dual port memories allow simultaneous read and write operations. Also, some special features are also configurable, like support for FIFO implementations with reduced utilization of fabric logic.

BRAMs are important for those designs with extensive use of on-chip memory. For those with low requirements in terms of memory, BRAMs are wasted and so its silicon area. Improving area efficiency of unused BRAMs has been a topic of research (Wilton, S. J. E., - 2002) by using memory to implement logic.

Medium-grained dedicated computational blocks are also present in FPGAs. A former FPGA considering dedicated computational blocks was the Xilinx Virtex FPGA (Xilinx, 2005). The block is an 18×18 signed multiplier mapped side by side with a BRAM. In some FPGAs (Altera 2006), the multiplier is configurable, that is, one 36×36 multiplier can be separated in eight smaller 9×9 multipliers. These dedicated multipliers were later replaced by DSP (Digital Signal Processing) modules with extra functionalities, like multiply-accumulate, pattern detection and pre-adders.

Some digital systems require a microprocessor. Since an FPGA is hardware programmable, one or more microprocessors can be programmed as any other digital circuit. To improve performance, some FPGAs include a hard microprocessor that can interface with the reconfigurable logic. One of the first approaches of this kind of FPGAs included an embedded processor connected to reconfigurable logic through a bus (Triscend, 2001). One year later, Altera integrated an ARM core in the Apex 20K family of FPGAs (Altera, 2002) and Xilinx introduced the Virtex II Pro family of FPGAs with the IBM PowerPC core embedded in the logic fabric of the FPGA (Xilinx, 2007). Integrating one or more processors in an FPGA opens many issues related to how the processor interfaces with the configurable logic. Initially, the processor was just another dedicated hardware block, but recent FPGAs follow a different approach considering the microprocessor as a central unit of the system and the programmable logic is used to implement accelerators to help the processor running tasks.

This vast heterogeneity of FPGA blocks increases exponentially the design space of FPGAs with many options for the designer. FPGA vendors have reduced this design space by introducing families of FPGAs according to the target applications. When targeting embedded processing, a microprocessor is fundamental, while high-performance computing requires more dedicated arithmetic blocks, digital signal processing requires arithmetic blocks and memory, and complex digital logic systems require more basic configurable logic blocks. Considering these target applications and markets, different families of FPGAs have different ratios of configurable logic blocks and dedicated units.

Recent FPGAs went a step further by integrating in the same device the programmable logic with a complete processing system with dual and quad-core processors, cache, on-chip memory and several more peripherals and interfaces to peripheral outside the FPGA forming a complete SoC (System-on-Chip) on FPGA. Examples of SoC FPGAs are the ZYNQ All Programmable SoC FPGA (Xilinx, 2018) with a 1.0 GHz (highest speed grade) dual-core ARM Cortex-A9 processor embedded within the programmable logic fabric. Arria V FPGA from Altera also includes an 800 MHz dual-core ARM Cortex-A9 MPCore integrated with the programmable logic. The Atmel FPSLIC is another FPGA device that uses a processor in combination with the programmable logic architecture of Atmel.

Application Domains of FPGAs

The evolution of FPGAs was driven not only by the process technology but also by target applications.

The very first expansion of FPGA production was mainly due to the communications market. Vendors of communication equipment use FPGAs for packet processing in switches and routers (Gibb et al., 2008). The communications area was constantly changing with new routing and switching algorithms. ASIC development processes could not keep up with the dynamism of the communications area, like FPGAs could. Bandwidth consumption is doubling every 2-3 years and communication requirements are everywhere from automotive to industrial equipment and consumer electronics. 400 Gbps (Giga bits per second) are here. FPGAs are used in the high-backplane interconnections and inside cellular networks to support 4G and 5G. The consequence was that vendors did customize their FPGAs for best performance

in the communications area with the inclusion of high-speed transceivers, with hundreds of dedicated multiply-accumulate blocks with a configurable pipeline capable to process packets at gigabit speed.

Another potential application for FPGAs is associated with data centers. Power requirements of data centers increase every year. Data center providers are balancing high-performance and power heat. FPGAs can work as accelerators of CPU clusters for a diverse processing tasks like image search, video transcode, image processing, data analytics, etc. (Putnam et al., 2014). FPGAs run many high-performance computing applications with efficient energy consumption with the advantage that can be reconfigured to run new algorithms. Reconfigurable DSPs play an important role to improve energy efficiency of high-performance reconfigurable computing. This aspect is more effective when fixed-point arithmetic can be used instead of floating-point since DSPs are more efficient to implement fixed-point arithmetic. Intel Arria 10 (Intel, 2017) was the first FPGA to include hardened single precision floating-point digital signal processing blocks.

FPGAs can also be used in some IoT (Internet of Things) embedded systems but only when high processing is required for encryption, data processing, etc. Otherwise, microcontroller devices are less expensive and less power hungry. Industrial IoT is another edge application of FPGAs to process motion control, machine-to-machine communication, real-time processing, etc.

Advanced Driver Assistance Systems (ADAS) can also take advantage of FPGAs for real-time analytics (object detection, object recognition and classification, pedestrian detection, among others).

Military and aerospace benefit from the reconfigurability and rapid development of FPGA-based systems to keep up with the very dynamic development of new standards in these areas. In space, the reconfigurability can be used to overcome design faults and new system functionality.

FUTURE RESEARCH DIRECTIONS

Integrated circuit process technology keep improving which increases logic density per chip. Fine-grained FPGAs is probably the most appropriate architecture for bit manipulation. However, the increasing silicon density of FPGAs permits the implementation of large complex systems, many of them operating on words of bits. The migration from bit to word processing moves the granularity of reconfigurable devices towards coarse grain architectures, where the atomic logic units may be arithmetic logic units or other complex structure. Coarse-grain reconfigurable computing is a research and development area with about three decades of existence with early works like PipeRench (Goldstein, et al., 2000), CHESS (Marshall et al., 1999), or RaPiD (Ebeling, 1996) with 16-bit multipliers, memories and arithmetic units. In some cases, the logic blocks are complete microprocessors (Taylor et al., 2002) (Buts et al., 2007).

In spite of all the research investment and development of coarse grain architectures, it is not clear how these architectures will achieve wide adoption. However, one think is clear, the fine-grain architectures of FPGAs may not be able to keep pace with the silicon density increase. New and more efficient coarse grain reconfigurable architectures are expected in the near future.

FPGA architectures will be greatly influenced by top emerging markets. Internet of Things (IoT) and edge computing encompasses a vast list of applications from smart consumer electronic to data analytics analysis in the edge. Future systems must deal with many different interfaces, heterogeneous computing requirements, low power consumption, etc. FPGAs are one possible solution for these new emerging systems. Other application market where FPGAs are already used is the autonomous driving. Several demanding real-time image processing tasks are essential for a complete autonomous driving experience.

Power consumption is a major concern of future FPGAs. Both cloud and edge computing are facing the dark silicon era, where transistors cannot be all powered together due to thermal power constraints. Power gating techniques must be applied not only to logic circuits but also to routing resources.

The rise of artificial intelligence and machine learning both in cloud and edge computing requires fast and heterogeneous architectures with high performance per watt. Xilinx is already planning future FPGA devices to deal with this market. The adaptive compute acceleration platform – ACAP (Xilinx, 2018) combines scalar engines, programmable engines and intelligent engines targeting data centers, 5G wireless and automotive driver assist.

CONCLUSION

The article describes the architecture of field programmable gate array devices, where logic routing and input/output can be configured for any particular digital system. FPGAs started as devices for glue-logic implementation and are now completely autonomous system-on-chip devices.

With the advance of the integrated circuit technology, its density have increased dramatically permitting the implementation of many different applications in different areas like consumer, industrial, automotive, high-performance and communications. Recent system-on-chip FPGA families have one or more microprocessor cores, memory, cache and reconfigurable logic to implement hardware/software systems in a single programmable device.

The reconfigurable logic of an FPGA consists of heterogeneous fine and coarse-grain blocks interconnected in an island-style routing architecture. The type and size of these heterogeneous blocks has being object of intensive research and discussion to find the most appropriate ones and the most appropriate balance.

The architecture of future FPGAs will be determined by energy consumption metrics and by the necessities of emergent markets, like IoT and machine learning.

Therefore, in spite of 35 years of existence, FPGAs are still an actual topic of intense research.

REFERENCES

- Aggarwal, A., & Lewis, D. (1994). Routing architectures for hierarchical field programmable gate arrays. *IEEE International Conference on Computer Design*, 475–478. 10.1109/ICCD.1994.331954
- Ahmed, E., & Rose, J. (2000). The effect of LUT and cluster size on deep-submicron FPGA performance and density. In *Proceedings of the 2000 ACM/SIGDA Eighth International Symposium on Field Programmable Gate Arrays*, (pp. 3–12). ACM Press. 10.1145/329166.329171
- Ahmed, E., & Rose, J. (2004). The effect of LUT and cluster size on deep-submicron FPGA performance and density. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 12(3), 288–298.
- Altera Corporation. (2002). *Excalibur device overview DS-EXCARM-2.0*. Author.
- Altera Corporation. (2003). *FLEX 10K embedded programmable logic device family, DS-F10K-4.2*. <http://www.altera.com/literature/ds/dsf10k.pdf>
- Altera Corporation. (2005). *Stratix II device handbook SII5V1-3.1*. [http://www.altera.com/literature/hb/stx2/stratix2 handbook.pdf](http://www.altera.com/literature/hb/stx2/stratix2%20handbook.pdf)

Altera Corporation. (2006). *Stratix III device handbook, ver 1.0*. http://www.altera.com/literature/hb/stx3/stratix3_handbook.pdf

Betz, V., & Rose, J. (1999). FPGA routing architecture: Segmentation and buffering to optimize speed and density. *Proceeding: ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, 140–149. 10.1145/296399.296428

Birkner, J., Chan, A., Chua, H. T., Chao, A., Gordon, K., Kleinman, B., Kolze, P., & Wong, R. (1992). A very-high-speed field-programmable gate array using metal-to-metal antifuse programmable elements. *Microelectronics Journal*, 23(7), 561–568. doi:10.1016/0026-2692(92)90067-B

Brown, S. D., Francis, R., Rose, J., & Vranesic, Z. (1992). *Field-Programmable Gate Arrays*. Kluwer Academic Publishers. doi:10.1007/978-1-4615-3572-0

Butts, M., Jones, A. M., & Wasson, P. (2007). A structural object programming model, architecture, chip and tools for reconfigurable computing. *Field-Programmable Custom Computing Machines, 2007. 15th Annual IEEE Symposium on. FCCM 2007*, 55–64. 10.1109/FCCM.2007.7

Carter, W., Duong, K., Freeman, R. H., Hsieh, H., Ja, J. Y., Mahoney, J. E., Ngo, L. T., & Sze, S. L. (1986). A user programmable reconfiguration gate array. *Proceedings of the IEEE Custom Integrated Circuits Conference*, 233–235.

Ebeling, C., Cronquist, D. C., & Franklin, P. (1996). RaPiD—reconfigurable pipelined datapath. In *6th International Workshop on Field-Programmable Logic and Applications*, (pp. 126–135). Springer.

El Gamal, A., Greene, J., Reyneri, J., Rogoyski, E., El-Ayat, K. A., & Mohsen, A. (1989). An architecture for electrically configurable gate arrays. *IEEE Journal of Solid-State Circuits*, 24(2), 394–398. doi:10.1109/4.18600

Gibb, G., Lockwood, J. W., Naous, J., Hartke, P., & McKeown, N. (2008). NetFPGA: An openplatform for teaching how to build gigabit-ratenetwork switches and routers. *IEEE Journal of Education*, 51(3), 364–369.

Goldstein, S. C., Schmit, H., Budiu, M., Cadambi, S., Moe, M., & Taylor, R. R. (2000). PipeRench: A reconfigurable architecture and compiler. *Computer*, 33(4), 70–77. doi:10.1109/2.839324

Guterman, D. C., Rimawi, L. H., Chiu, T.-L., Halvorson, R. D., & McElroy, D. J. (1979). An electrically alterable nonvolatile memory cell using a floating-gate structure. *IEEE Transactions on Electron Devices*, 26(4), 576–586. doi:10.1109/T-ED.1979.19462

Intel. (2017). *Intel®Arria® 10 Native Floating-Point DSP Intel® FPGA IP User Guide*. Intel.

Intel. (2019). https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/hb/stratix-10/s10_datasheet.pdf

Marple, D., & Cooke, L. (1992). An MPGA compatible FPGA architecture. *Proceedings of the IEEE Custom Integrated Circuits Conference*, 4.2.1–4.2.4.

Marshall, A., Stansfield, T., Kostarnov, I., Vuillemin, J., & Hutchings, B. (1999). A reconfigurable arithmetic array for multimedia applications. *ACM/SIGDA Seventh International Symposium on Field Programmable Gate Arrays*, 135–143. 10.1145/296399.296444

- Masud, M. I., & Wilton, S. (1999). A new switch block for segmented FPGAs. *Proceedings: International Workshop on Field Programmable Logic and Applications*, 274–281. 10.1007/978-3-540-48302-1_28
- Putnam, A., Caulfield, A. M., Chung, E. S., Chiou, D., Constantinides, K., Demme, J., Esmaeilzadeh, H., Fowers, J., Gray, J., Haselman, M., Hauck, S., Heil, S., Hormati, A., Kim, J.-Y., Lanka, S., Larus, J. R., Peterson, E., Prashanth, G., Smith, A., . . . Burger, D. (2014). A Re-configurable Fabric for Accelerating Large-Scale Datacenter Services. *International Symposium on Computer Architecture (ISCA)*.
- Plessey Semiconductor. (1989). *ERA60100 preliminary data sheet*. Author.
- Taylor, J., Kim, J., Miller, J., Wentzlaff, D., Ghodrat, F., Greenwald, B., Hoffman, H., Johnson, P., Jae-Wook Lee, Lee, W., Ma, A., Saraf, A., Seneski, M., Shnidman, N., Strumpfen, V., Frank, M., Amarasinghe, S., & Agarwal, A. (2002). The Raw microprocessor: A computational fabric for software circuits and general purpose programs. *IEEE Micro*, 22(2), 25–35. doi:10.1109/MM.2002.997877
- Triscend Corporation. (2001). *Triscend E5 configurable system-on-chip platform product description TCH300-0001-001*. <http://www.keil.com/dd/docs/datashts/triscend/te5xx.pdf>
- Wahlstrom, S. E. (1967). Programmable logic arrays — cheaper by the millions. *Electronics (Basel)*, 40, 90–95.
- Wilton, S. J. E. (2002). Implementing logic in FPGA memory arrays: Heterogeneous memory architectures. *IEEE International Conference on Field-Programmable Technology*, 142–149. 10.1109/FPT.2002.1188675
- Xilinx. (2005a). *Virtex-4 family overview*. DS112(v1.4). <http://direct.xilinx.com/bvdocs/publications/ds112.pdf>
- Xilinx. (2005b). *Virtex-II platform FPGAs: Complete data sheet*. DS031(v3.4). <http://direct.xilinx.com/bvdocs/publications/ds031.pdf>
- Xilinx. (2007). *Virtex-II Pro and Virtex-II Pro X Platform FPGAs: Complete Data Sheet*. DS083(v4.6). <http://direct.xilinx.com/bvdocs/publications/ds083.pdf>
- Xilinx. (2018). *Versal: The First Adaptive Compute Acceleration Platform (ACAP)*. White paper, WP505.
- Xilinx Inc. (2018). *Zynq-7000 SoC Data Sheet: Overview*. Author.

ADDITIONAL READING

- Brown, S. (1996). FPGA architectural research: A survey. *IEEE Design & Test of Computers*, 13(4), 9–15. doi:10.1109/54.544531
- Farooq, Umer & Marrakchi, Zied & Mehrez, Habib. (2012). FPGA architectures: An overview. . doi:10.1007/978-1-4614-3594-5_2
- Kuon, I., Tessier, R., & Rose, J. (2007). FPGA Architecture: Survey and Challenges. *Foundations and Trends in Electronic Design Automation*, 2(2), 135–253. doi:10.1561/10000000005
- Parab, J. S., Gad, R. S., & Naik, G. M. (2018). *Hands-on Experience with Altera FPGA Development Boards*. Springer. doi:10.1007/978-81-322-3769-3

Shawahna, A., Sait, S. M., & El Maleh, A. (2019). FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review. *IEEE Access: Practical Innovations, Open Solutions*, 7, 7823–7859. doi:10.1109/ACCESS.2018.2890150

Vanderbauwhede, W., & Benkrid, K. (2013). *High-Performance Computing Using FPGAs*. Springer. doi:10.1007/978-1-4614-1791-0

KEY TERMS AND DEFINITIONS

Block RAM (BRAM): A random access memory embedded in FPGAs for data storage. BRAMs can be interconnected to implement larger RAMs.

Coarse-Grained Granularity: Refers to the granular size of reconfigurable architectures consisting of arrays of units reconfigurable at the word level. A typical example is an arithmetic unit.

Configurable Logic Block (CLB): The basic configurable logic resource of an FPGA. CLBs can be interconnected to implement complex logic functions.

Digital Signal Processing Block (DSP): Configurable hard arithmetic blocks in FPGAs for arithmetic operations. DSPs can be interconnected to implement larger arithmetic operations.

Field-Programmable Gate Array (FPGA): An integrated circuit whose hardware can be configured after manufacturing.

Fine-Grained Granularity: Refers to the granular size of the reconfigurable hardware at the bit level. Fine-grained granularity is typically associated to field-programmable gate arrays.

Granularity: Refers to the relative size of the reconfigurable elements in a reconfigurable hardware architecture.

Hierarchical Routing: An interconnection architecture where routing organizes the FPGA in clusters of logic blocks. Neighbor blocks are interconnected with local wires, while logic blocks in different clusters must go through one or more higher levels of routing.

Island-Style Routing: An interconnection architecture where routing organizes the FPGA as a two dimensional flat array surrounded by routing channels.

Look-Up Table (LUT): A small memory used to implement functions in an FPGA.

Reconfigurable Computing (RC): A computing paradigm that uses reconfigurable hardware for computing purposes.

Reconfigurable Hardware: A hardware structure whose logic elements and their interconnections can be reconfigured to implement a particular logic circuit.

Design of a Simple and Low-Cost Calculator in the Laboratory Using FPGA

2

Debapriya Mukherjee*University of Calcutta, India***Kaustav Das***University of Calcutta, India***Arpita Das** <https://orcid.org/0000-0002-5939-2382>*University of Calcutta, India*

INTRODUCTION

It is a common practice to realize specific operations/algorithms in embedded systems using programmable general-purpose processors. However, an alternative solution to software implementation is the design of application specific dedicated hardware in order to achieve high performance system at relatively low cost. The advancement in VLSI technology makes it attractive for implementation in any application specific task to minimize the size and power requirements which in turn reduce the cost.

In this view, we aim to design a prototype of a simple, low cost calculator in the laboratory which would enable us to perform the basic arithmetic operations like addition, subtraction, multiplication, division along with some logical operations. Though a great amount of work had been done on Arithmetic and Logic Unit (ALU) design (O-Cisneros et al. 2005; William 2006; Xiao et al. 2008; Oztekin et al. 2011; Singh 2014), it is still a challenging task to implement the algorithms successfully in hardware. FPGAs are matrix of configurable logic blocks connected by programmable interconnects and are used to implement desired functionality after manufacturing (Brown et al. 1992; Rose et al. 1993). Due to the programmable nature of FPGAs, they are used in many application fields, such as ASIC prototyping, wireless communications, video/image processing, surveillance, industrial automation, information appliances, home networking & monitoring, residential set top boxes etc. Beside the promising computational platform of FPGAs, parallel nature of computing and overall non recurring costs make them significantly faster for some applications (Rose et al. 1993; Oldfield & Dorf 1995).

The objective of this work is to design an alternative, simple, low power, cost effective calculator in the laboratory. FPGA prototyping of this reusable, dynamic calculator is a reliable way to verify the design in hardware (Lysaght 1993; Oldfield & Dorf 1995; Kalte et al, 2002). Initial step of design is to frame out the simple arithmetic and logical operations and then it would be upgraded for scientific calculations (inclusion of trigonometric operations and logarithmic calculations). In this work, design guideline of simple arithmetic circuits such as adders, subtractors, multiplier, divider and some logical units are studied and assembled to configure the dynamic calculator. The inputs may be applied either

DOI: 10.4018/978-1-7998-3479-3.ch021

by using switches in the FPGA board or a conventional PS/2 keyboard that is connected externally but interfaced to FPGA and the results are displayed on the (2 × 16) paneled LCD screen available on the FPGA board. The design algorithm of this operation is very simple and hardware resource requirement is also minimal. Hence any low-level FPGA hardware is sufficient to implement the design in the laboratory.

Background Architecture of FPGA

Since 1960, there have already been four generations of ICs .viz SSI (small scale integration), MSI (medium scale integration), LSI (large scale integration), and VLSI (very large scale integration). An electronic circuit usually consists of a CPU, ROM, RAM and other peripherals on one board. VLSI is a process of combining thousands of transistors into a single chip occupying a relatively smaller area (<https://electronicsforu.com/resources/learn-electronics/vlsi-developments-ic-fabrication>).

The primary advantages of using VLSI technology are:

- Reducing the size of circuits.
- Reducing the effective cost of the devices.
- Increasing the operating speed of circuits.
- Requirement of less power than discrete components.

Field Programmable Gate Arrays (FPGAs) are semiconductor devices that may be programmed according to specific tasks after manufacturing, while Application Specific Integrated Circuits (ASICs) are manufactured for specific design tasks. There are some specific software to configure FPGA as simple as an AND gate to as complex as the multi-processor (Zhu et al. 2009). The FPGA stores the specific configuration in RAM, which is why the configuration is lost when there is no power connectivity (Lysaght 1993; Oldfield & Dorf 1995). Hence, they must be configured every time the power is on. FPGAs are prefabricated on silicon chips that can be programmed electrically to implement the desired functions. As shown in Fig. 1, FPGA Architecture consists of three major components —

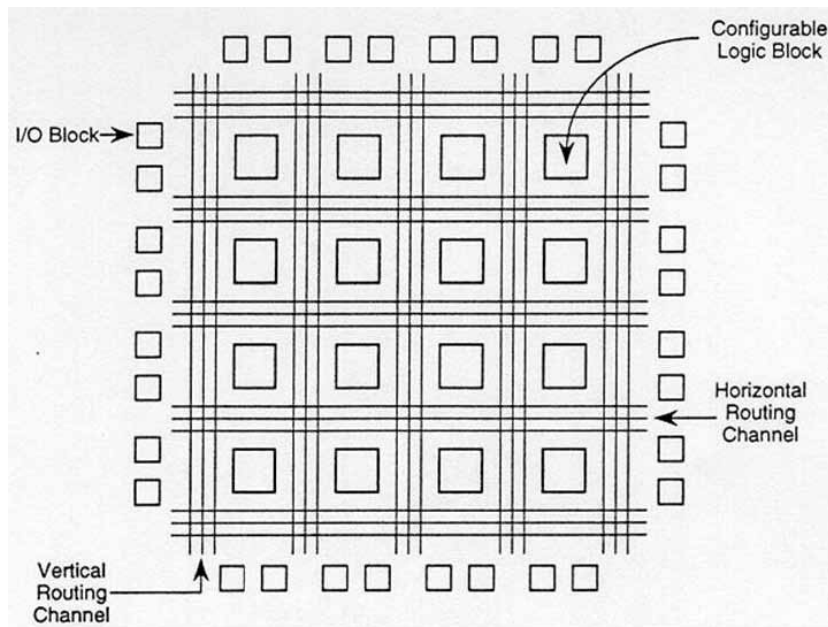
- Configurable Logic Blocks (CLB), which implement logic functions
- Programmable Routing (interconnects), which implements the functions
- I/O blocks, which are used to make off-chip connections

Configurable Logic Blocks: These blocks provide basic computation and storage elements used in digital systems. A basic logic element consists of programmable look-up-tables, a flip-flop, and some fast carry logic to reduce the area and delay cost.

Programmable Routing: The programmable routing establishes a connection between logic blocks and Input/Output blocks to complete a user-defined design unit. It consists of horizontal and vertical routing channels which again containing multiplexers, pass transistors and tri-state buffers. Pass transistors and multiplexers are used in a logic cluster to connect the logic elements.

Programmable I/O Blocks: The programmable I/O pads are used to interface the logic blocks and routing architecture to the external components. The I/O pads and the surrounding logic circuits form the I/O blocks.

Figure 1. General architecture of Xilinx FPGA



Motivation of the Work

With the rising demand for portable, power efficient and reconfigurable design of electronic circuits and systems in the laboratory, we intend to design a calculator using FPGA which is capable of performing simple arithmetic operations as well as the logical operations and it can preserve the facilities for future modifications that might be necessary. For this purpose we will need to design the arithmetic and logic unit first, followed by the sub modules like multiplexer, binary to BCD converter, interfacing module for LCD display each of which has a definite function and the combined operations of these units will finally make our attempt to use the FPGA as a prototype of dynamic calculator successful.

In this work, the software that is used is *Xilinx ISE 14.2* and the language for designing each of the sub modules is *VHDL*, a hardware description language used in electronic design (Bhasker 1999; IEEE stand. Manual 1993; Perry 2002).

Proposed Architecture

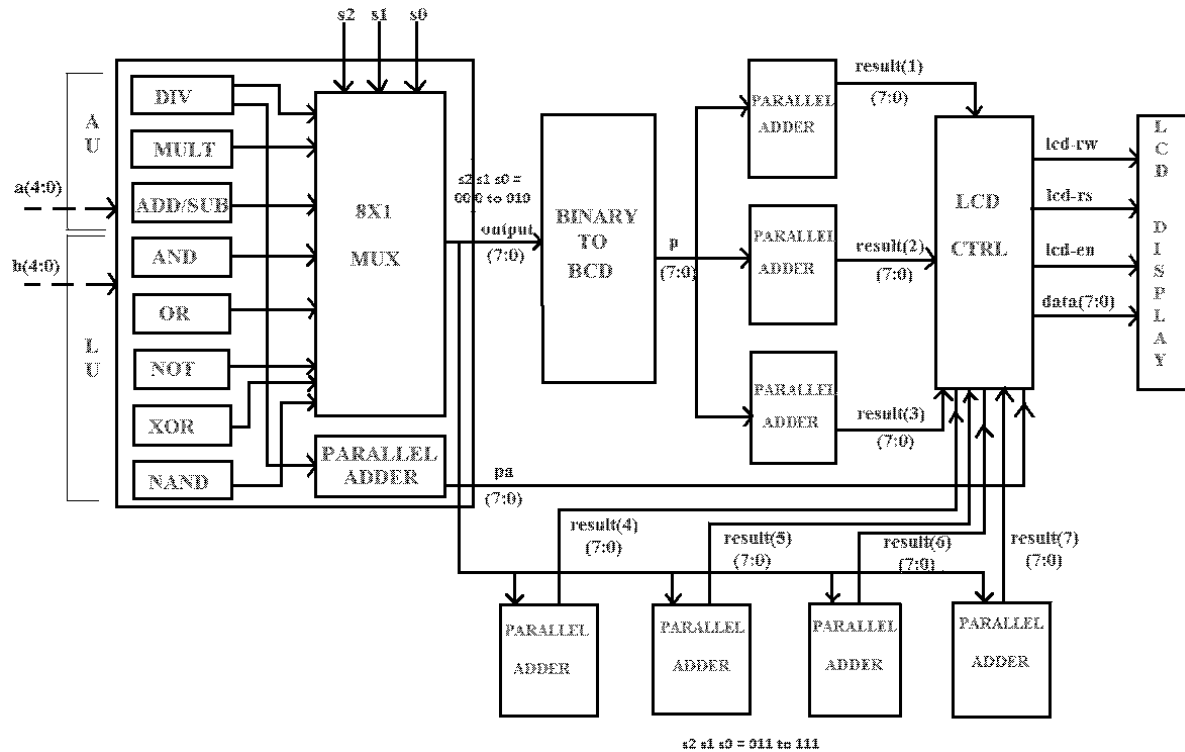
To begin with the initial design of dynamic calculator using FPGA, we need to describe the overall architecture of the calculator as shown in Figure 2. –

According to the block diagram of Figure 2, operations of the proposed ALU architecture can be described as follows.

Whenever the data inputs $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$ are applied by the user and the operations are specified by the *selection lines* of 8×1 MUX, the output of that particular operation will be computed and finally displayed on the LCD screen.

In the proposed architecture, the choice of arithmetic/logical operations is made by the combinations of select lines $(s_2 s_1 s_0)$ of MUX and hence the variation (000 to 010) is restricted for arithmetic operations and the rest of the combinations are considered for different logical operations. Moreover, the output

Figure 2. Block Diagram of the proposed ALU and LCD interfacing architecture



of arithmetic operations need binary to BCD conversion to generate the decimal equivalent value and then corresponding ASCII codes of LCD screen are dynamically produced by the 8-bit parallel adders. However, the results of logical operations have no need for BCD conversion and thus they are directly pipelined to 8-bit parallel adders for generation of ASCII codes.

In the following section, we have described three fundamental sub-modules of the proposed architecture and their functionalities in details.

- Arithmetic Unit (AU):** It is the very fundamental building block of a calculator and performs the following basic arithmetic operations on the operands provided by the user.
 - DIV**– It is a division unit which will divide an 8-bit numerator by a 4-bit denominator to generate a 4-bit quotient and a 3-bit remainder.
 - MULT** – It is a multiplier unit which will multiply a 4-bit multiplicand and a 4-bit multiplier and will generate a product of 8 bits.
 - ADD/SUB** – It is an adder/subtractor unit which will either add or subtract according to the mode value. Operands are two binary numbers each of 4 bits and finally result will be generated in the signed data format.
- Logic unit (LU):** It is another fundamental building block of the calculator and our proposed architecture consists of all basic logical operations as follows -
 - AND** – It is the first logical unit which will perform the AND operation between two binary numbers $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$.

- **OR** – It is the second logical unit which will perform the OR operation between two binary numbers $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$.
- **NOT** – It is the third logical unit which will perform the inversion/negation operation on any 4-bit binary number.
- **XOR** – It is the fourth logical unit which will perform the XOR operation between two binary numbers $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$.
- **NAND** – It is the fifth logical unit which will perform the NAND operation between two binary numbers $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$.
- **Additional Sub-modules:** Few additional building blocks are required to represent the desired output of the prototype in the LCD display panel of FPGA board.
 - A **Multiplexer** unit is used to decide which of the eight operations will be computed when the inputs are applied.
 - The output of the multiplexer will either go to the **Binary to BCD converter** or directly propagate to the **8-bit parallel adder** unit depending on the **select line combination** of the multiplexer. Binary to BCD conversion is required for generating the corresponding BCD form of the binary outputs of arithmetic operations whereas the results of logical operations are directly pipelined to next stages, since there is no need for BCD conversion in those cases.
 - The output of BCD converter/direct results of logical operations will propagate to the inputs of the **8-bit parallel adder** circuits in order to generate the ASCII codes of LCD panel for displaying the results. In this architecture, each of the parallel adder is used to display **single digit** on the LCD screen. Since the ASCII codes for the numerical digits of LCD panel follows a particular pattern like - (0 @30H, 1@31H, 2@32H, 3@33H, 4@34H, 5@35H, 6@36H, 7@37H, 8@38H, 9@39H), one of the input of each 8-bit parallel adder may be kept fixed at 30H (00110000) whereas the other input will remain variable. Therefore, every time the parallel adder will add the variable input with 30H (00110000) to generate the required ASCII codes.
 - The outputs of all the parallel adders will then be interfaced to the LCD module of FPGA board by establishing the proper control signals of controller unit.

In this way we may assemble the individual blocks to display the results on the LCD screen. Further modifications in the design block can be made to upgrade the calculator for more versatile uses. We have described the detailed architecture and functionality of each of the blocks in the following sections.

4.1 Adder/Subtractor: It is a fundamental building block of an arithmetic unit in dynamic calculator. The operations of the proposed ripple carry adder for adding/subtracting two signed binary number is described in the following sections.

Working Principle: As shown in Figure 3, let us assume one of the 4-bit numbers to be $a(a_3 a_2 a_1 a_0)$ and the other be $b(b_3 b_2 b_1 b_0)$. There are 2 separate block of adder circuit each of which consists of a set of 4 full adders. Also there are 2 sets of XOR gates that remain associated with the 2 adder blocks. The mode signal M decides whether the operation to be executed is Addition or Subtraction. If $M=0$, addition will be performed on the 2 inputs and if $M=1$ then subtraction will be performed on the same inputs. The 4-bit sum of the 1st adder block is assumed to be $s'(s'_3 s'_2 s'_1 s'_0)$ and the 2nd adder block as $s(s_3 s_2 s_1 s_0)$.

The algorithmic steps for addition operation (Ray-Chaudhuri 2009; Lewin 1974) are given below:

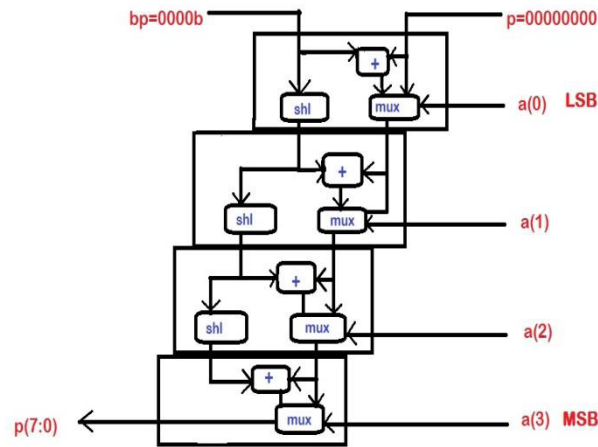
- In the first block of the circuit one of the inputs a ($a_3 a_2 a_1 a_0$) enters the full adders directly whereas the other input b ($b_3 b_2 b_1 b_0$) enters a set of four 2 input XOR gates. The mode signal M acts as the 2nd input for each of the XOR gates.
- The output of the XOR gates remain same as the input when $M=0$ i.e. during addition. Therefore b enters the 1st adder block unchanged.
- The input carry C_{in} is same as M for addition.
- In each of the full adder of the 1st block a_i gets added to b_i and the intermediate carry passes on until the output carry C_{out} is obtained.
- C_{out} enters another XOR gate named as XI along with M and the final output carry is generated which remains same as C_{out} since $M=0$.
- The **SUM** is the output of the 1st adder block i.e. s' ($s'_3 s'_2 s'_1 s'_0$).

- In the first block of the circuit one of the inputs a ($a_3 a_2 a_1 a_0$) enters the full adders directly whereas the other input b ($b_3 b_2 b_1 b_0$) enters into a set of four number of 2 input XOR gates. The mode signal M acts as the 2nd input for each of the XOR gates.
- Since $M = 1$ during subtraction, the XOR gates generate the 1's complement of the input b and when this is added to $C_{in} = 1$ gives the 2's complement of b . This is added to a in the first block.
- Now, If $a > b$ then the difference is generated by simple addition operation as described in the previous section. If $a < b$ then the following steps have to be computed.
 $\emptyset$ M enters an XOR gate named $X1$ along with C_{out} of the 1st block as well as an AND gate named $A1$. The other input of $A1$ is the output of $X1$. Together they generate the sign bit as the output of $A1$.

- Ø In the 2nd block of the circuit the sum bits from the previous block, $s'(s_3' s_2' s_1' s_0')$ enter as one of the inputs of another set of four XOR gates. The other input of the 2 input XOR gates is the sign bit which is 1 if $a < b$.
- Ø This generates the 1's complement of the inputs which is further added to C_{in} to generate the 2's complement. $C_{in} = 1$ since it is same as the output of AI .
- Ø The 2's complement of $s'(s_3' s_2' s_1' s_0')$ is added to $a(a_3 a_2 a_1 a_0)$ which has been kept 0000 to produce the final output $s(s_3 s_2 s_1 s_0)$. This is the **difference** of a and b .

4.2 Multiplier: Multiplier is a widely used important arithmetic unit. Figure 4 describes the multiplication operation of two binary coded positive integer numbers. This is also the core unit for signed binary multiplication.

Figure 4. Block diagram of a Multiplier circuit



Working Principle: As shown in Figure 4, let us assume that the multiplier is $a(a_3 a_2 a_1 a_0)$ and the multiplicand is $b(b_3 b_2 b_1 b_0)$. Also $b_p = 0000b_3 b_2 b_1 b_0$ and initially the product is $p = 00000000$. We have considered a variable p_v which is the sum of p_v and b_p such that,

$$p_v = p_v + b_p \quad (1)$$

The algorithmic steps of the proposed multiplier (Ray-Chaudhuri 2009; Shirriff 2013) are given below:

- Since the multiplicand and the multiplier are of 4 bits so shifting will take place *thrice* and addition will take place *four* times.
- Initially $p_v = p_v + b_p = 0000b_3 b_2 b_1 b_0$. This is done by the first adder circuit symbolized as '+' as shown in the figure.
- The content of b_p is shifted left by 1 bit by the shifting block 'shl'. Now $b_p = 000b_3 b_2 b_1 b_0$.
- The output of the first adder i.e, the content of p_v enters into the 2nd adder circuit by a 2:1 multiplexer if its select input $a(0)=1$, otherwise the previous value of p_v will go through the multiplexer directly.
- The output of the shifting block i.e, the current content of b_p also enters the 2nd adder circuit.

- Now b_p gets added to the current value of p_v .
- Once again b_p gets shifted by 1 bit to the left by 2nd shifting block 'shl'. Now $b_p = 00b_3b_2b_1b_000$.
- If $a(1) = 1$ the output of the adder moves on to the 3rd adder circuit via a 2:1 multiplexer whose select input is $a(1)$, otherwise once again the previous value of p_v will pass on to the next adder.
- The shifted contents of b_p enters into the 3rd adder circuit as its 2nd input. After addition the contents of p_v gets changed.
- Once again the contents of b_p get shifted by the 3rd shifting block 'shl' so that $b_p = 0b_3b_2b_1b_0000$.
- If $a(2) = 0$ the previous value of p_v passes on directly to the 4th and last adder without any change, otherwise the current value of p_v will move on to the last adder via a 2:1 multiplexer whose select input is $a(2)$.
- Ultimately either the output of the 4th adder circuit or the previous content of p_v will be obtained as the final output which is the PRODUCT depending upon whether the select input of the 4th multiplexer $a(3) = 1$ or $a(3) = 0$.

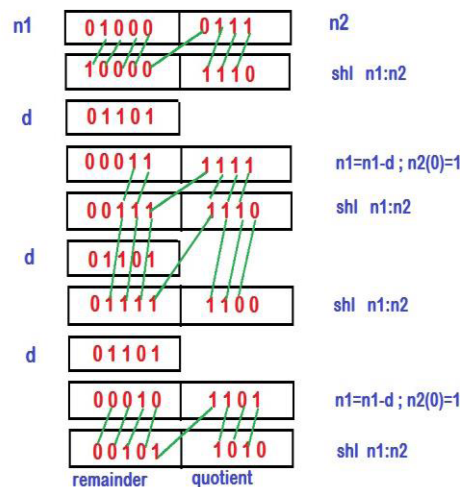
4.3 Divider: Division is also an important arithmetic operation in the design of dynamic calculator. Figure 5 describes the division operation of two unsigned binary numbers.

Working Principle: Let us consider the numerator to be a 8-bit number (n) and the denominator to be a 4-bit number (d). As shown in the diagram, the numerator has been divided into two parts, n_1 and n_2 . n_1 is a 5-bit number that consists of the 4 bits from the MSB with a 0 inserted at the MSB position and n_2 is a 4 bit number which consists of the remaining bits of the numerator upto the LSB. d is also made a 5 bit number by inserting a 0 at its MSB position. This is done to keep n_1 and d of equal bits. The contents of n_1 and n_2 are shifted left by 1 bit, and if $n_1 > d$ then

$$n_1 = n_1 - d ; n_2(0) = 1 \quad (2)$$

Also we have a provision of displaying the word 'ERROR' if the denominator value is 0.

Figure 5. Block diagram of a Division circuit



The algorithmic steps of the divider unit (Sites 1974; Srinivasan, 1994) are given below:

- The contents of n_1 and n_2 are shifted left by 1 bit and as a result 0 enters as LSB of n_2 .
- If $n_1 > d$ then $n_1 = n_1 - d$ and $n_2(0) = 1$. If $n_1 < d$ then n_1 and n_2 remain unchanged.
- The process of shifting gets repeated thrice. Everytime either $n_1 = n_1 - d$ or both n_1 and n_2 remain unaltered depending upon whether $n_1 > d$ or $n_1 < d$.
- After the 4th and last shift n_2 holds the **quotient** and n_1 holds the **remainder**.

4.4 AND/OR/XOR/NAND: These are few fundamental bitwise logical operations which are used as the fundamental building blocks of digital logic circuits (Katz 1994; Lala 1996).

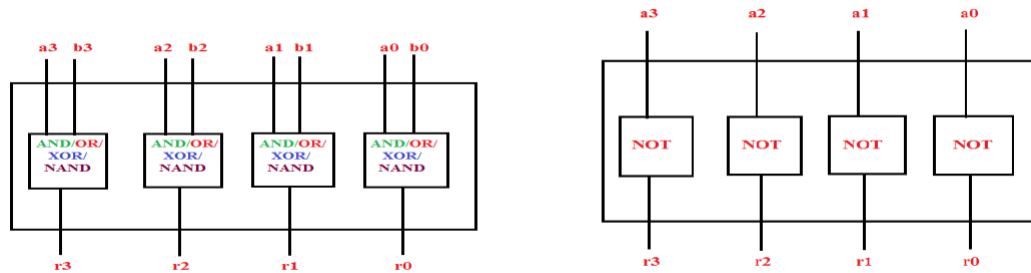
Working Principle: The above logic unit takes two inputs as $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$ and then performs 4-bit AND/OR/XOR/NAND operations for producing 4-bit output $r(r_3 r_2 r_1 r_0)$.

4.5 Inversion: A logic inverter or NOT gate performs a basic logical operation which inverts its binary input (Lala 1996; Enderton 2001).

Figure 6 Show the pictorial presentation of 4-bit logical AND/OR/XOR/NAND and Inversion unit.

Working Principle: As shown in Figure 6, the above inverter block performs 4-bit inversion operation on input $a(a_3 a_2 a_1 a_0)$ and produces a 4-bit output $r(r_3 r_2 r_1 r_0)$.

Figure 6. Pictorial illustration of 4-bit logical AND/OR/XOR/NAND as well as Inverter Block



4.6 8:1 MULTIPLEXER: Multiplexer is a circuit arrangement that selects one of several inputs and let it propagate to the output, hence it is also known as data selector (Ray-Chaudhuri 2009). In our work, 8×1 MUX is designed in order to select among three arithmetic and five logical operations. Three select line combinations determine which operation is to be performed when the data inputs are applied by the user. Accordingly the result of any one of the said operation is sent by the MUX to the following section of the module.

The algorithmic steps of 8×1 MUX are given below:

- When selection-line $(S_2 S_1 S_0) = 000$, the output of **Division** operation will be selected and the quotient will propagate through the MUX whereas the remainder will enter an 8 bit parallel adder which will be discussed in the following part of the report.
- When selection-line $(S_2 S_1 S_0) = 001$, the output of **Multiplication** operation will be selected and the product will propagate through the MUX to its output.
- When selection-line $(S_2 S_1 S_0) = 010$, the output of **Add/Sub** operation will be selected and the Sum/Difference will propagate through the MUX to its output.

- When selection-line $(S_2 S_1 S_0) = 011$, the output of the **AND** gates will propagate through the MUX to its output.
- When selection-line $(S_2 S_1 S_0) = 100$, the output of the **OR** gates will propagate through the MUX to its output.
- When selection-line $(S_2 S_1 S_0) = 101$, the output of the **NOT** gates will propagate through the MUX to its output.
- When selection-line $(S_2 S_1 S_0) = 110$, the output of the **XOR** gates will propagate through the MUX to its output.
- When selection-line $(S_2 S_1 S_0) = 111$, the output of the **NAND** gates will propagate through the MUX to its output.

The output of the 8×1 MUX either enters as the input of the Binary-to-BCD converter which again converts 8-bit binary result to its corresponding BCD having a maximum of 10 bits or directly pipelined to 8-bit parallel adders depending upon the value of select lines of the MUX.

4.7 Binary to BCD Converter: A binary to BCD converter will convert an 8 bit binary input to its BCD equivalent. For this conversion the method of Shifting and Adding is implemented with some intermediate corrections made during this process (Ray-Chaudhuri 2009; Lewin 1974). The corrections are required due to two reasons:

- (a) Any 4 bit group exceeding 1001 upto 1111 does not represent correct BCD form. This needs a correction when it occurs within a decade due to the left shift.
- (b) The second correction is needed when a bit from the MSB position of a decade is shifted to the LSB position of the next higher decade adjacent to it

All the necessary corrections take place automatically if 0011(=3) is added to the decade which contain a number greater than 0100(=4). The value of any decade may exceed 0100 just after 3 shifts or at the end of 4 shifts. The following algorithm is called “**Add and Shift Algorithm**”.

Figure 7 describes the block diagram of an 8-bit binary to BCD converter.

The algorithmic steps of binary to BCD converter are given below:

- i) Place the binary number outside and to the right of the spaces of decade-0.
- ii) Shift left one bit at a time until the decade value becomes greater than 4 (0100).
- iii) When the decade value is found to be greater than 0100, add 0011 to it and shift left immediately.
- iv) Repeat step (iii) until the LSB of the given binary number has been shifted into the least significant position of decade-0.
- v) The process stops after the last shift and the content of the decade is the required BCD.

4.8 8-BIT PARALLEL ADDER: An 8-bit parallel adder is used to generate each of the ASCII codes corresponding to its input. Once the 8-bit binary result of any one of the arithmetic or logical operation is converted to its BCD format, it is sent to the parallel adder so that it can produce an 8-bit ASCII code for each decimal digit which is to be displayed on the LCD screen. Figure 8 shows the conventional diagram of an 8-bit parallel adder. Since we chose to display the results in the LCD screen by using 3 digits, we need to implement 3 parallel adder circuits in order to generate the ASCII codes of the 3 digits separately.

The algorithmic steps of the proposed 8-bit parallel adder are given below:

Figure 7. Block diagram of Binary to BCD converter

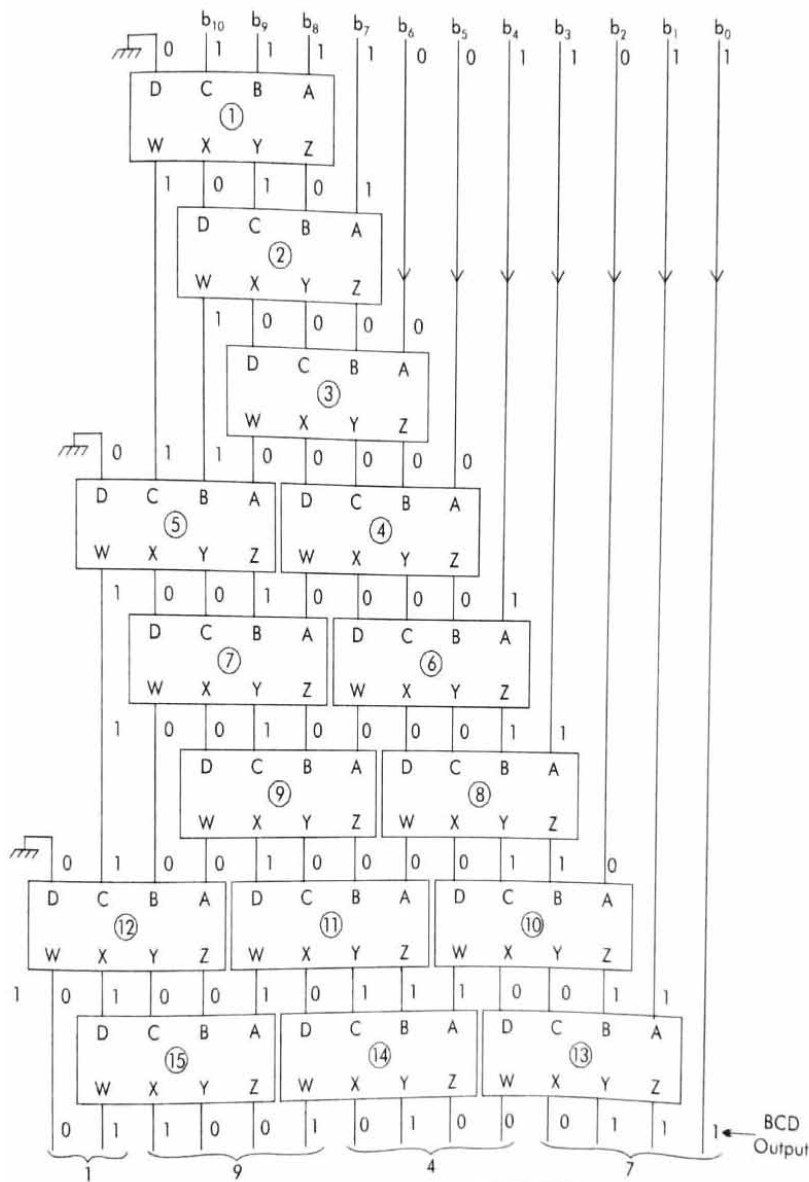
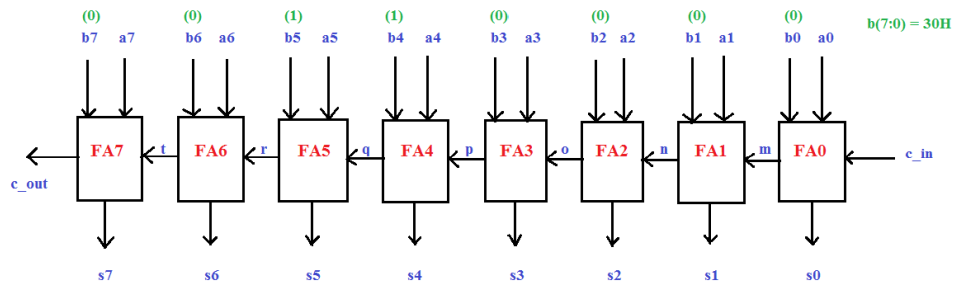


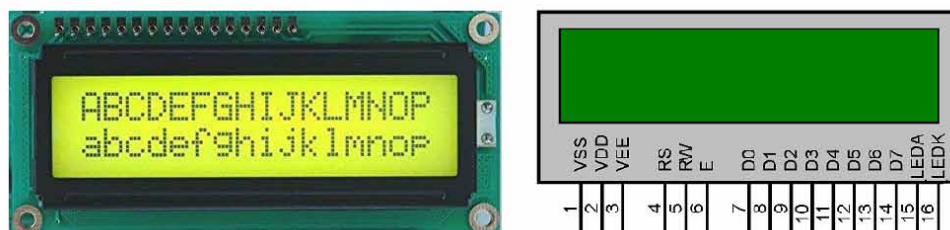
Figure 8. 8 bit parallel adder



- The 8-bit binary parallel adder has two inputs, a ($a_7 a_6 a_5 a_4 a_3 a_2 a_1 a_0$) and b ($b_7 b_6 b_5 b_4 b_3 b_2 b_1 b_0$). The block consists of 8 full adder components. Each full adder adds two input bits along with the carry bit which is forwarded through the carry output of the previous full adder.
- Since in the LCD board, ASCII codes of the numerical digits of Xilinx Spartan-3E FPGA board are: ['0' -- 30H ; '1' -- 31H ; '2' -- 32H ; '3' -- 33H ; '4' -- 34H ; '5' -- 35H ; '6' -- 36H ; '7' -- 37H ; '8' -- 38H ; '9' -- 39H]; the input ' b ' is kept fixed at 30H (00110000) whereas the input ' a ' remains variable. Therefore every time each parallel adder adds the input ' a ' with 30H to generate the required ASCII codes for displaying the results (Xilinx User Guide, 2016; VHDL Designer's Guide).

4.9 LCD Interfacing for Display: A liquid crystal display or LCD draws its definition from its name itself. It is the combination of two states of matter, the solid and the liquid. LCD uses a liquid crystal to produce a visible image. Liquid crystal displays are super-thin technology display screen that are generally used in laptop computer screen, TVs, mobile phones, portable video games and also in FPGA board as shown in Figure 9. LCD's technologies allow displays to be much thinner way in compare to cathode ray tube technology. Proper interfacing procedure of LCD will help to display the output results into (2 × 16) panel as shown in Figure 9.

Figure 9. (i) (2 × 16) panel LCD display; (ii) Pin Configurations



Pin Configurations and the connections between LCD screen and FPGA chip for interfacing are described in detail in Table 1. LCD panel has 8-bit bidirectional data bus interface with some control lines such as Enable signal (E), Register Select (RS) and Read/Write signal (R/W) (Xilinx User Guide, 2016; web link of LCD interfacing tutorials)

The operations of three control signals are given below:

- The *Register Select* control line enables the kit to identify the input as either a Command (to set the cursor, to introduce blank spaces etc.) or a Data (the characters that are to be displayed). $RS = 1$ implies that input is data and $RS = 0$ implies that input is a command.
- The *Read/Write* control enables the kit to either perform read operation or write operation. ($R/W = 1$ implies read operation and $R/W = 0$ implies write operation)
- If the *Enable* input is high, i.e. $E = 1$, then only the data gets latched to the data bus.

Hence by assembling all these modules, we have attempted to design the prototype of a simple calculator that computes and displays the results of the above mentioned arithmetic/logical operations on the LCD screen of the FPGA board.

Table 1. Configurations of LCD board

Pin No	Function	Name
1	Ground (0V)	Ground
2	Supply voltage; 5V (4.7V – 5.3V)	V _{cc}
3	Contrast adjustment; through a variable resistor	V _{EE}
4	Selects command register when low; and data register when high	Register Select
5	Low to write to the register; High to read from the register	Read/write
6	Sends data to data pins when a high to low pulse is given	Enable
7	8-bit data pins	DB0
8		DB1
9		DB2
10		DB3
11		DB4
12		DB5
13		DB6
14		DB7
15	Backlight V _{CC} (5V)	Led+
16	Backlight Ground (0V)	Led-

Experimental Results

In the proposed architecture of dynamic calculator, we are applying the inputs with the switches available on the Spartan-3E series of FPGA board. However, there are 16 switches assigned for the inputs, among which three switches are assigned for three select lines of 8×1 MUX and one is for Mode selection of ADD/SUB unit in our design, hence the number of inputs that can be applied as data inputs a and b , gets restricted to 12. Hence with these 12 number of available input switches, four bit data operations are considered in this proposed architecture such as $a(a_3 a_2 a_1 a_0)$ and $b(b_3 b_2 b_1 b_0)$. The output results computed by the calculator are displayed in the (2 × 16) panel LCD screen of the FPGA board.

In the following sections, experimental results of some arithmetic and logical operations are shown with descriptions.

- (i) Result of Addition: When the MUX is selected for addition operation ($S_2 S_1 S_0 = 010$, and $M=0$), if we provide the inputs as –

$$a(a_3 a_2 a_1 a_0) = \mathbf{0111} \text{ (decimal equivalent 7),}$$

$$b(b_3 b_2 b_1 b_0) = \mathbf{0100} \text{ (decimal equivalent 4),}$$

Output result will be **1011** (decimal equivalent **11**), which is shown in the LCD screen of Figure 10.

Figure 10. Addition Result



- (ii) Subtraction results: When the MUX is selected for subtraction operation ($S_2S_1S_0 = 010$, and $M=1$), if we provide the inputs as –

$$a (a_3 a_2 a_1 a_0) = \mathbf{0010} \text{ (decimal equivalent 2),}$$

$$b (b_3 b_2 b_1 b_0) = \mathbf{0101} \text{ (decimal equivalent 5),}$$

Output result will be **1011**(decimal equivalent **-3**), which is shown in the LCD screen of Figure 11.

Figure 11. Subtraction Result



- (iii) Multiplication result: When the MUX is selected for multiplication operation ($S_2S_1S_0 = 001$), if we provide the inputs as –

$$a(a_3a_2a_1a_0) = \mathbf{1100} \text{ (decimal equivalent 12),}$$

$$b(b_3b_2b_1b_0) = \mathbf{0011} \text{ (decimal equivalent 3),}$$

Output result will be **100100** (decimal equivalent **36**), which is shown in the LCD screen of Figure 12.

Figure 12. Multiplication Result



- (iv) Division results: When the MUX is selected for division operation ($S_2S_1S_0 = 000$), if we provide the inputs as –

$$a(a_3a_2a_1a_0) = \mathbf{0111} \text{ (decimal equivalent 7),}$$

$$b(b_3b_2b_1b_0) = \mathbf{0100} \text{ (decimal equivalent 4),}$$

The quotient will be **0001**(decimal equivalent **1**); remainder will be **0011** (decimal equivalent **3**); which are shown in the LCD screen of Figure 13.

Again for another division operation, if we provide the inputs as –

$$a(a_3a_2a_1a_0) = \mathbf{0001} \text{ (decimal equivalent 1),}$$

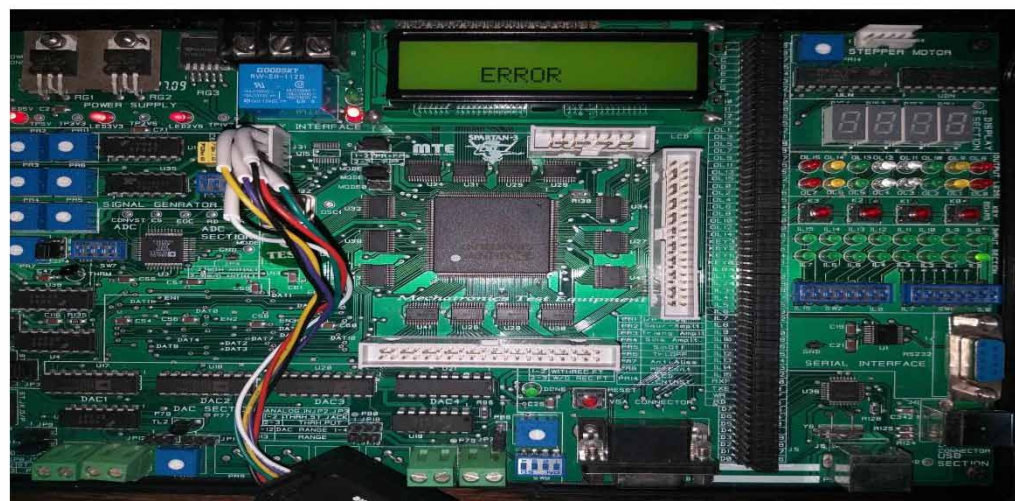
$$b(b_3b_2b_1b_0) = \mathbf{0000} \text{ (decimal equivalent 0),}$$

The output result will be “**Error**”, which is shown in the LCD screen of Figure 14.

Figure 13. Division Result



Figure 14. Division Result



Discussions

In this work, we have attempted to provide a sample overview of FPGA prototyping in the laboratory setup. Since the proposed architecture of calculator computes only some simple arithmetic and logical operations, this may be implemented in any low-level starter kit of FPGA. Presently the proposed architecture utilizes only 7% of CLB slices and 19% of I/O resources of Spartan-3E FPGA board and the design is optimized for speed.

In this proposed architecture, due to the availability of restricted number of input switches, data operations are also bounded by 4-bits. This restriction/drawback may be overcome by interfacing the standard PS-2 keyboard with the FPGA kit and then the architecture will be more flexible for general purpose

use. Moreover, another major drawback of this ALU prototype is computation of only unsigned binary numbers for arithmetic operations which will also be upgraded to both signed and unsigned operations for more versatile use.

Moreover, in this prototype of ALU, we have designed a ripple carry adder for simplicity which can be replaced by carry-look ahead approach for achieving the better speed of computation. A carry-look ahead adder (<https://www.electronicshub.org/carry-look-ahead-adder/>) is a fast parallel adder as it reduces the propagation delay by more complex hardware. It is designed by transforming the ripple-carry adder circuit such that the carry logic of the adder is changed into two-level logic and the circuit at any stage does not have to wait for the generation of carry-bit from the previous stage. Hence, the carry bit can be evaluated at any instant of time.

FUTURE DIRECTIONS

In this study, we only focused on the design and implementation of simple non-scientific arithmetic operations in a low cost manner. A complete design cycle from feeding of input data for any of four basic arithmetic operations to visualization of results in conventional decimal form is presented in this work. Design of this prototype is very easy to understand for the trainee level and it is the main strength of the present work.

All other scientific calculations such as trigonometric, logarithmic and power-law computations may be achieved by exploiting the CORDIC algorithm (<https://www.allaboutcircuits.com/technical-articles/an-introduction-to-the-cordic-algorithm/>). CORDIC is a hardware-efficient iterative method which uses rotations of input vectors by a scaling factor to calculate a wide range of elementary functions. The list of the functions that can be calculated from this rotation is relatively long such as arc-tan, arc-sin, arc-cos, hyperbolic and logarithmic functions, polar to rectangular transform, Cartesian to polar transform etc. Moreover, multiplication and division operations can also be obtained from the variants of rotation.

Furthermore, the prototype will take a more sophisticated look if some provision might be kept for simple mathematical equation solving capability in the architecture.

ACKNOWLEDGMENT

This work is partially supported by the Center of Excellence (CoE) in Systems Biology and Biomedical Engineering, University of Calcutta funded by the World Bank, MHRD India.

REFERENCES

- An introduction to CORDIC algorithm. (n.d.). <https://www.allaboutcircuits.com/technical-articles/an-introduction-to-the-cordic-algorithm/>
- Bhasker, J. (1999). *A VHDL Primer* (3rd ed.). Pearson Education Publication.
- Brown, S., Francis, R., Rose, J., & Vranesic, Z. (1992). *Field-Programmable Gate Arrays*. Kluwer Academic Publishers. doi:10.1007/978-1-4615-3572-0
- Carry look ahead adder. (n.d.). <https://www.electronicshub.org/carry-look-ahead-adder/>

- Enderton Herbert. (2001). *A Mathematical Introduction to Logic* (2nd ed.). Academic Press.
- IEEE Standard VHDL Language Reference Manual. (1993). *Std 1076-1993*. IEEE.
- Intoduction to VLSI technology. (n.d.). <https://electronicsforu.com/resources/learn-electronics/vlsi-developments-ic-fabrication>
- Kalte, H., Langen, D., Vonnahme, E., Brinkmann, A., & Rückert, U. (2002). Dynamically Reconfigurable System-on-Programmable-Chip. *Proc. of the 10th Euromicro Workshop on Parallel, Distributed and Network-based Processing (PDP 2002)*, 234-242.
- Katz, R. H. (1994). *Contemporary Logic Design*. The Benjamin/Cummings Publishing Company.
- Lala, P. K. (1996). *Practical Digital Logic Design and Testing*. Prentice Hall.
- LCD interfacing tutorials. (n.d.). <https://www.pantechsolutions.net/fpga-tutorials/16x2-lcd-interfacing-with-spartan3e-fpga-development-kit>
- Lewin Douglas. (1974). *Logical Design of Switching Circuits*. Academic Press.
- Lysaght, P., & Dunlop, J. (1993). Dynamic Reconfiguration of Field Programmable Gate Arrays. *Proceedings of the 3rd Int. Workshop on Field Programmable Logic and Applications*.
- O-Cisneros, S., R-Panduro, J.J., S Muro, J., & Boemo, E. (2005). Rapid prototyping of a Self-Timed ALU with FPGAs. *Proc. of Int. Conf. on Reconfigurable Computing and FPGAs (ReConFig 2005)*.
- Oldfield, J., & Dorf, R. (1995). *Field Programmable Gate Arrays*. John Wiley & Sons.
- Oztekin, H., Temurtas, F., & Gulbag, A. (2011). A Modular Approach to Arithmetic and Logic Unit Design on a Reconfigurable Hardware Platform for Educational Purpose. In *Intelligent Computing and Information Science. ICICIS 2011. Communications in Computer and Information Science*, (vol. 135). Springer. doi:10.1007/978-3-642-18134-4_55
- Perry Douglas, L. (2002). *VHDL Programming by Example* (4th ed.). McGraw-Hill New York.
- Ray Chaudhuri, D. (2009). *Digital Circuits, Volume-1* (2nd ed.). Platinum Publishers.
- Rose, J., El Gamal, A., & Sangiovanni-Vincentelli, A. (1993). Architecture of Field-Programmable Gate Arrays. *Proceedings of the IEEE*, 81(7), 1013–1029. doi:10.1109/5.231340
- Shirriff Ken. (2013). *The Z-80 has a 4-bit ALU. Here's how it works*. Academic Press.
- Singh, H. (2014). Design of enhanced arithmetic logical unit for hardware genetic processor. *Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI)*. 10.1109/ICACCI.2014.6968248
- Sites, R. L. (1974). Serial binary division by ten. *IEEE Transactions on Computers*, C-23(12), 1299–1301. doi:10.1109/T-C.1974.223849
- Srinivasan, P. (1994). Constant-division algorithms. *IEE Proc. Comput. Tech.*, 141(6), 334–340.
- VHDL Designer's Guide. (n.d.). https://www.doulos.com/knowhow/vhdl_designers_guide/models/binary_bcd/
- William, S. (2006). *Computer Organization & Architecture: Designing for Performance* (7th ed.). Pearson Prentice Hall.

Xiao, T., & Liu, F. (2008). 16-bit teaching microprocessor design and application. *IEEE International Symposium on IT in Medicine and Education, IEEE Conferences*, 160–163.

Xilinx UG230 Spartan-3E FPGA Starter Kit Board User Guide. (2016). Xilinx.

Zhu, Y., Weng, T., & Keng, C. (2009). Enhancing Learning Effectiveness in Digital Design Courses Through the Use of Programmable Logic. *IEEE Transactions on Education*, 52(1), 151–156. doi:10.1109/TE.2008.921796

KEY TERMS AND DEFINITIONS

8×1 MUX: Multiplexer with 8 input data-line and single output terminal.

ALU: Arithmetic and logic unit.

ASCII Codes: American standard code for information interchange.

ASIC: Application-specific integrated circuit.

BCD Conversion: Binary coded decimal conversion.

CPU: Central processing unit.

FPGA: Field programmable gate array.

I/O Block: Input/output block.

LCD: Liquid crystal display.

MSB/LSB: Most significant bit/least significant bit.

RAM: Random access memory.

ROM: Read only memory.

VHDL: Very high-speed integrated circuit hardware description language.

VLSI: Very large-scale integration.

XOR Gates: Exclusive-OR gate.

Modeling of Uncertain Nonlinear System With Z-Numbers

Raheleh Jafari

 <https://orcid.org/0000-0001-7298-2363>

School of Design, University of Leeds, Leeds, UK

Sina Razvarz

 <https://orcid.org/0000-0003-1549-7307>

Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico

Alexander Gegov

 <https://orcid.org/0000-0002-6166-296X>

School of Computing, University of Portsmouth, UK

Satyam Paul

 <https://orcid.org/0000-0003-4720-0897>

Department of Engineering Design and Mathematics, University of the West of England, Bristol, UK

INTRODUCTION

Interpolation technique is extensively implemented in order to function estimation (Boffi & Gastaldi, 2006; Jafarian et al., 2016; Mastlylo, 2010). In (Schroeder et al., 1991) a systolic algorithm in order to interpolate as well as evaluate of polynomials by utilizing a linear array of processors is proposed. Two-dimensional polynomial interpolation is described in (Zoli, 2008). In (Neidinger, 2009) a multivariable interpolation method is proposed. In (Olver, 2006), the multivariate Vandermode matrix is implemented. Recently, Smooth function estimation is extensively used (Szabados & Vertesi 1990; Tikhomirov, 1990) which leads a model by using Lagrange interpolating polynomials at the points of product grids (Barthelmann et al., 2000; Xiu & Hesthaven, 2005). However, if there are uncertainties in the interpolation points, none of these techniques will work properly.

The fuzzy equation is taken to be the general form of the fuzzy polynomial. The concept of fuzzy modeling is based on finding the fuzzy coefficients of the fuzzy equation. Various methods have been developed (Goetschel & Voxman, 1986; Kajani et al., 2005; Mamdani, 1976; Mazandarani & Kamyad, 2013; Salahshour et al., 2012; Takagi & Sugeno, 1985; Wang & liu, 2011; Jafari & Yu, 2015a, 2015b, 2015c, 2015d, 2017a; Jafari et al, 2019a, 2016; Razvarz & Jafari, 2017a, 2017b; Razvarz et al., 2017, 2018). In (Friedman et al., 1998) a general fuzzy linear system is studied utilizing the embedding method. In (Buckley & Qu, 1990) the necessary and sufficient conditions for linear as well as quadratic equations are presented in order to contain a solution in a case that the parameters are either real or complex fuzzy numbers. The homotypic analysis technique is discussed in (Abbasbandy, 2006). The Newton's method is proposed in (Abbasbandy & Ezzati, 2006b). In (Allahviranloo et al., 2007), fixed point technique is suggested in order to solve a system of fuzzy nonlinear equations. Nevertheless, these techniques are so much complex.

DOI: 10.4018/978-1-7998-3479-3.ch022

Iterative technique (Llanas & Sainz, 2006), interpolation technique (Waziri & Majid, 2012) and Runge-Kutta technique (Pederson & Sambandham, 2008) can be applied for obtaining the numerical solution of the fuzzy equation. These techniques can also be applied to fuzzy differential equations (Kajani & Lupulescu, 2009). Neural networks can also be applied in order to find the solutions of the fuzzy equations (Cybenko, 1989; Ito, 2001; Jafari & Razvarz, 2017, 2018, 2015e; Jafari & Yu, 2017b, 2017c; Jafari et al., 2018a, 2018b, 2019b; Jafarian et al., 2016; Jafarian & Jafari, 2012; Yu & Jafari, 2019). In (Buckley & Eslami, 1997), neural network technique is utilized in order to solve the fuzzy quadratic equation. In (Jafari & Razvarz, 2017; Jafarian et al., 2012), the neural network is used in order to obtain the numerical solution of the dual fuzzy equation. In (Mosleh, 2013) neural network method is utilized in order to find the approximate solution of the fully fuzzy matrix equation.

In this book chapter, the neural network technique is utilized in order to estimate the Z-number coefficients of fuzzy equations. The concept of Z-number is used in various areas related to the analysis of the decisions (Kang et al., 2012a; Zadeh, 2006). Z-number is more precise when compared with the fuzzy number. Fewer studies are reported concerned to the topic of Z-numbers (Gardashova, 2014). Converting Z-number into the usual fuzzy set is discussed in (Kang et al, 2012b). In (Aliiev et al., 2015) the theoretical basics of the arithmetic of discrete Z-numbers are presented.

BACKGROUND

This work is the first attempt in finding the Z-number coefficients of fuzzy equations. In order to train the neural network, the backpropagation approach is used. The work is organized as follows. In Section 2, some basic definitions related to the Z-numbers are given. The proposed method for obtaining the Z-number coefficients of the fuzzy equations is demonstrated in Section 3. Some examples with applications in mechanics is given in Section 4. Section 5 concludes the work and provides discussions on further work.

Nonlinear System Modeling With Fuzzy Equations and Z-Numbers

Provide A common discrete-time nonlinear system is depicted as

$$\vartheta_{r+1} = f[\vartheta_r, q_r], \quad w_r = g[\vartheta_r] \quad (1)$$

where $q_r \in \mathbb{R}^n$ is the input vector, $\vartheta_r \in \mathbb{R}^l$ is an internal state vector, also $w_r \in \mathbb{R}^m$ is the output vector. f as well as g are generalized nonlinear smooth functions $f, g \in C^\infty$. Define $W_r = [w_{r+1}^T, w_r^T, \dots]^T$ as well as $Q_r = [q_{r+1}^T, q_r^T, \dots]^T$. Assume $\frac{\partial w}{\partial \vartheta}$ is non-singular at the instance $\vartheta_r = 0$, $Q_r = 0$, so the following model is extracted

$$w_r = Y[w_{r-1}^T, w_{r-2}^T, \dots, q_r^T, q_{r-1}^T, \dots] \quad (2)$$

in which $Y(\bullet)$ is a nonlinear difference equation representing the plant dynamics, q_r as well as w_r are calculable scalar input and output respectively.

The nonlinear systems explained in (2), can be written as the below mentioned linear-in-parameter model

$$w_r = \sum_{i=1}^{\gamma} a_i f_i(\vartheta_r) \quad (3)$$

where a_i is the linear parameter, $f_i(\vartheta_r)$ is nonlinear functions. The variables of this function are quantifying input and output.

The uncertain nonlinear systems can be modeled using the linear-in-parameter models with uncertain parameters. In this book chapter it has been assumed that the model of the nonlinear systems (3) contains uncertainties in the a_i , ϑ_r as well as v_r . These uncertainties have been stated in the form of Z-numbers (Zadeh, 2011).

Definition 1. Suppose c is:

- 1) normal, there is $\varsigma_0 \in \mathfrak{R}$ where $c(\varsigma_0) = 1$,
- 2) convex, $c(\beta\varsigma + (1-\beta)\sigma) \geq \min\{c(\varsigma), c(\sigma)\}$, $\forall \varsigma, \sigma \in \mathfrak{R}, \forall \beta \in [0, 1]$,
- 3) upper semi-continuous on $\mathfrak{R}$, $c(\varsigma) \leq c(\varsigma_0) + \varepsilon$, $\forall \varsigma \in N(\varsigma_0)$, $\forall \varsigma_0 \in \mathfrak{R}$, $\forall \varepsilon > 0$, $N(\varsigma_0)$ is a neighborhood,
- 4) $c^+ = \{\varsigma \in \mathfrak{R}, c(\varsigma) > 0\}$ is compact, so c is a fuzzy variable, $c \in E : \mathfrak{R} \rightarrow [0, 1]$.

The fuzzy variable c is demonstrated as

$$c = (\underline{c}, \overline{c}) \quad (4)$$

in which $\underline{c}$ is the lower-bound variable also, $\overline{c}$ is the upper-bound variable.

Definition 2. The z -number is made up of two components $Z = [c(\varsigma), p]$. The first component $c(\varsigma)$ is the restriction on a real-valued uncertain variable ς . The second component p is a measure of reliability of c . p can be reliability, strength of belief, probability or possibility. The z -number can be stated as Z^+ -number, in a case $c(\varsigma)$ be a fuzzy number also p be the probability distribution of ς . If $c(\varsigma)$ as well as p are fuzzy numbers, the z -number can be stated as Z^- -number.

The Z^+ -number contains more information when compared with the Z^- -number. In this book chapter, the definition of Z^+ -number is utilized, i.e., $Z = [c, p]$, c is a fuzzy number also, p is a probability distribution.

The most common membership functions which define the fuzzy numbers are the triangular function

$$\mu_{\varsigma} = H(s, u, v) = \begin{cases} \frac{\varsigma-s}{u-s} & s \leq \varsigma \leq u \\ \frac{v-\varsigma}{v-u} & u \leq \varsigma \leq v \end{cases} \text{ otherwise } \mu_{\varsigma} = 0 \quad (5)$$

and trapezoidal function

$$\mu_{\varsigma} = H(s, u, v, w) = \begin{cases} \frac{\varsigma-s}{u-s} & s \leq \varsigma \leq u \\ \frac{w-\varsigma}{w-v} & v \leq \varsigma \leq w \text{ otherwise } \mu_{\varsigma} = 0 \\ 1 & u \leq \varsigma \leq v \end{cases} \quad (6)$$

The probability measure can be stated as

$$P = \int_{\mathcal{R}} \mu_c(\varsigma) p(\varsigma) d\varsigma \quad (7)$$

where p is the probability density of ς also, $\mathcal{R}$ is the restriction on p . For discrete z -numbers we have

$$P(c) = \sum_{i=1}^{\gamma} \mu_c(\varsigma_i) p(\varsigma_i) \quad (8)$$

Definition 3. The α -level for fuzzy number c is stated as

$$[c]^\alpha = \{\varsigma \in \mathcal{R} : c(\varsigma) \geq \alpha\} \quad (9)$$

where $0 < \alpha \leq 1$, $c \in E$

Therefore $[c]^0 = c^+ = \{\varsigma \in \mathcal{R}, c(\varsigma) > 0\}$. As $\alpha \in [0, 1]$, $[c]^\alpha$ is bounded, $\underline{c}^\alpha \leq [c]^\alpha \leq \bar{c}^\alpha$. The α -level of c between $\underline{c}^\alpha$ and $\bar{c}^\alpha$ can be defined as

$$[c]^\alpha = (\underline{c}^\alpha, \bar{c}^\alpha) \quad (10)$$

$\underline{c}^\alpha$ as well as $\bar{c}^\alpha$ are the function of α . We define $\underline{c}^\alpha = d_A(\alpha)$, $\bar{c}^\alpha = d_B(\alpha)$, $\alpha \in [0, 1]$.

Definition 4. The α -level of the z -number $Z=(c, p)$ is defined as follows

$$[Z]^\alpha = ([c]^\alpha, [p]^\alpha) \quad (11)$$

where $0 < \alpha \leq 1$. $[p]^\alpha$ is computed by the Nguyen's theorem

$$[p]^\alpha = p([c]^\alpha) = p([\underline{c}^\alpha, \bar{c}^\alpha]) = [\underline{p}^\alpha, \bar{p}^\alpha] \quad (12)$$

where $p([c]^\alpha) = \{p(\varsigma) \mid \varsigma \in [c]^\alpha\}$. Therefore, $[Z]^\alpha$ is stated as

$$[Z]^\alpha = (\underline{Z}^\alpha, \bar{Z}^\alpha) = ((\underline{c}^\alpha, \underline{p}^\alpha), (\bar{c}^\alpha, \bar{p}^\alpha)) \quad (13)$$

where $\underline{p}^\alpha = \underline{c}^\alpha p(\underline{\varsigma}_i^\alpha)$, $\bar{p}^\alpha = \bar{c}^\alpha p(\bar{\varsigma}_i^\alpha)$, $[\varsigma_i]^\alpha = (\underline{\varsigma}_i^\alpha, \bar{\varsigma}_i^\alpha)$.

The same as with the fuzzy numbers (Jafari & Yu, 2015b), three main operations are defined for the z -numbers: $\oplus$, $\ominus$ and $\odot$, which indicate sum, subtract and multiply respectively. In this book chapter, the proposed operations are different from the ones in (Zadeh, 2005).

Suppose $Z_1=(c_1,p_1)$ as well as $Z_2=(c_2,p_2)$ are two discrete z -numbers expressing the uncertain variables ς_1 and ς_2 , so, $\sum_{k=1}^{\gamma}p_1(\varsigma_{1k})=1$, $\sum_{k=1}^{\gamma}p_2(\varsigma_{2k})=1$. The following operation is defined

$$Z_{12} = Z_1 * Z_2 = (c_1 * c_2, p_1 * p_2) \quad (14)$$

where $* \in \{\oplus, \ominus, \odot\}$.

The operations used for the fuzzy numbers are stated as (Jafari, 2015b),

$$[c_1 \oplus c_2]^\alpha = [\underline{c}_1^\alpha + \underline{c}_2^\alpha, \bar{c}_1^\alpha + \bar{c}_2^\alpha]$$

$$[c_1 \ominus c_2]^\alpha = [\underline{c}_1^\alpha - \underline{c}_2^\alpha, \bar{c}_1^\alpha - \bar{c}_2^\alpha] \quad (15)$$

$$[c_1 \odot c_2]^\alpha = \begin{pmatrix} \min \{ \underline{c}_1^\alpha \underline{c}_2^\alpha, \underline{c}_1^\alpha \bar{c}_2^\alpha, \bar{c}_1^\alpha \underline{c}_2^\alpha, \bar{c}_1^\alpha \bar{c}_2^\alpha \} \\ \max \{ \underline{c}_1^\alpha \underline{c}_2^\alpha, \underline{c}_1^\alpha \bar{c}_2^\alpha, \bar{c}_1^\alpha \underline{c}_2^\alpha, \bar{c}_1^\alpha \bar{c}_2^\alpha \} \end{pmatrix}$$

For the discrete probability distributions, the following relation is defined for all $p_1 * p_2$ operations

$$p_1 * p_2 = \sum_i p_1(\varsigma_{1,i}) p_2(\varsigma_{2,(\gamma-i)}) = p_{12}(\varsigma) \quad (16)$$

The Hukuhara difference is defined as (Aliev et al., 2016),

$$Z_1 \ominus_H Z_2 = Z_{12}$$

$$Z_1 = Z_2 \oplus Z_{12} \quad (17)$$

In a case that $Z_1 \ominus_H Z_2$ exists, the α -level can be defined as

$$[Z_1 \ominus_H Z_2]^\alpha = [\underline{Z}_1^\alpha - \underline{Z}_2^\alpha, \bar{Z}_1^\alpha - \bar{Z}_2^\alpha] \quad (18)$$

α Clearly, $Z_1 \ominus_H Z_1 = 0$, $Z_1 \ominus Z_1 \neq 0$.

Moreover, the generalized Hukuhara difference is defined as (Bede & stefanini, 2013),

$$Z_1 \ominus_{gH} Z_2 = Z_{12} \Leftrightarrow \begin{cases} 1) Z_1 = Z_2 \oplus Z_{12} \\ 2) Z_2 = Z_1 \oplus (-1)Z_{12} \end{cases} \quad (19)$$

By taking into consideration the α -level, we have

$$[Z_1 \ominus_H Z_2]^\alpha = [\min\{\underline{Z}_1^\alpha - \underline{Z}_2^\alpha, \bar{Z}_1^\alpha - \bar{Z}_2^\alpha\}, \max\{\underline{Z}_1^\alpha - \underline{Z}_2^\alpha, \bar{Z}_1^\alpha - \bar{Z}_2^\alpha\}]$$

and If also, if $Z_1 \ominus_H Z_2$ exists, $Z_1 \ominus_H Z_2 = Z_1 \ominus_{gH} Z_2$. The conditions for the existence of $Z_{12} = Z_1 \ominus_{gH} Z_2 \in E$ are

$$\begin{aligned} 1) & \begin{cases} \underline{Z}_{12}^\alpha = \underline{Z}_1^\alpha - \underline{Z}_2^\alpha \text{ and } \bar{Z}_{12}^\alpha = \bar{Z}_1^\alpha - \bar{Z}_2^\alpha \\ \text{with } \underline{Z}_{12}^\alpha \text{ increasing, } \bar{Z}_{12}^\alpha \text{ decreasing, } \underline{Z}_{12}^\alpha \leq \bar{Z}_{12}^\alpha \end{cases} \\ 2) & \begin{cases} \underline{Z}_{12}^\alpha = \bar{Z}_1^\alpha - \bar{Z}_2^\alpha \text{ and } \bar{Z}_{12}^\alpha = \underline{Z}_1^\alpha - \underline{Z}_2^\alpha \\ \text{with } \underline{Z}_{12}^\alpha \text{ increasing, } \bar{Z}_{12}^\alpha \text{ decreasing, } \underline{Z}_{12}^\alpha \leq \bar{Z}_{12}^\alpha \end{cases} \end{aligned} \quad (20)$$

where $\forall \alpha \in [0, 1]$.

Suppose c is a triangular function, the absolute value of the z -number $Z=(c,p)$ is defined as

$$|Z(\varsigma)| = (|s_1| + |u_1| + |v_1|, p(|s_2| + |u_2| + |v_2|)) \quad (21)$$

Let c_1 as well as c_2 are triangular functions, the supremum metric for z -numbers $Z_1=(c_1,p_1)$ and $Z_2=(c_2,p_2)$ is expressed as

$$D(Z_1, Z_2) = d(c_1, c_2) + d(p_1, p_2) \quad (22)$$

where $d(\bullet, \bullet)$ is the supremum metrics for fuzzy sets (Jafari, 2015b). $D(Z_1, Z_2)$ has the below mentioned properties,

$$\begin{aligned} D(Z_1 + Z, Z_2 + Z) &= D(Z_1, Z_2) \\ D(Z_2, Z_1) &= D(Z_1, Z_2) \\ D(bZ_1, kZ_2) &= |b| D(Z_1, Z_2) \\ D(Z_1, Z_2) &\leq D(Z_1, Z) + D(Z, Z_2) \end{aligned} \quad (23)$$

where $b \in \mathfrak{R}$, $Z=(c,p)$ is z -number, also c is triangle function.

Definition 5. Suppose $\hat{Z}$ is the space of z -numbers. The α -level of the z -number valued function $H : [0, s] \rightarrow \hat{Z}$ is defined as

$$H(c, \alpha) = [\underline{H}(c, \alpha), \bar{H}(c, \alpha)] \quad (24)$$

where $c \in \hat{Z}$, for each $\alpha \in [0, 1]$.

Using the definition of Generalized Hukuhara difference, the gH -derivative of H at c_0 is defined as

$$\frac{d}{dt} H(c_0) = \lim_{\zeta \rightarrow 0} \frac{1}{\zeta} [H(c_0 + \zeta) \ominus_{gH} H(c_0)] \quad (24)$$

In (25), $H(c_0 + \zeta)$ as well as $H(c_0)$ represent symmetric pattern with Z_1 and Z_2 respectively given in (19).

In this book chapter, the aim of the modeling is to find the Z-number coefficient a_i for the below fuzzy equation

$$w_r^* = \sum_{i=1}^{\gamma} a_i f_i(\vartheta_r) \quad (25)$$

in such a way that the output of the plant w_r approaches to the desired output w_r^* .

Definition 6. The distance between two Z-numbers, $\psi, \phi \in \tilde{Z}$, is described as the Hausdorff metric $D_H[\psi, \phi]$,

$$D_H[\psi, \phi] = \max \left\{ \sup_{(a_1, c_1) \in \psi} \inf_{(a_2, c_2) \in \phi} (d(a_1, a_2) + d(c_1, c_2)), \sup_{(a_1, c_1) \in \phi} \inf_{(a_2, c_2) \in \psi} (d(a_1, a_2) + d(c_1, c_2)) \right\} \quad (27)$$

$d(a, c)$ is the supremum metrics defined for fuzzy sets.

Lemma 1. Let $\varpi \subset \tilde{Z}$ is a compact set, in this case, ϖ is uniformly support-bounded, i.e. there exists a compact set $\Psi \subset \mathfrak{R}$, in such a way that $\forall \psi \in \varpi$,

$$\text{Supp}(\psi) \subset \Psi. \quad (28)$$

Lemma 2. Suppose $\psi, \phi \in \tilde{Z}$, also $\alpha \in (0, 1]$, in this case the following is concluded:

- (i) if $f, g : \mathfrak{R} \rightarrow \mathfrak{R}$ are continuous, $[f(\psi)]^\alpha = f([\psi^\alpha])$ holds;
- (ii) if $f : \mathfrak{R} \rightarrow \mathfrak{R}$ are continuous, so $f(\text{Supp}(\psi)) = \text{Supp}(f(\psi))$.

proof. As (i) concludes from (Yang & Gao, 1993), so just (ii) will be proved. Initially, $\overline{f(D_1)} = f(\overline{D_1})$ is shown for $D_1 \subset \mathfrak{R}$. As $f(D_1) \subset f(\overline{D_1})$, also is closed by the continuity of f , therefore, $\overline{f(D_1)} \subset f(\overline{D_1})$. Also, for randomly given $\theta \in f(\overline{D_1})$, there exists a sequence $\{a_n \mid n \in N\} \subset \mathfrak{R}$, $a \in \mathfrak{R}$, in such a way that $a_n \rightarrow a (n \rightarrow +\infty)$, $\theta = f(a)$. Since f is continues so, $\lim_{n \rightarrow +\infty} f(a_n) = f(a) = \theta$.

However, $f(a_n) \in f(D_1)$, therefore, $\theta \in f(D_1)$. So $\overline{f(D_1)} \subset f(\overline{D_1})$. Hence $\overline{f(D_1)} = f(\overline{D_1})$.

Since,

$$\begin{aligned} \text{Supp}(f(\Psi)) &= \overline{\{\theta \in \mathfrak{R} \mid (f(\Psi))(\theta) > 0\}} \\ f(\text{Supp}(\Psi)) &= f(\overline{\{a \in \mathfrak{R} \mid \Psi(a) > 0\}}) \end{aligned} \quad (29)$$

the following is concluded

$$\begin{aligned} f(\text{Supp}(\Psi)) &= \overline{f(\{a \in \mathfrak{R} \mid \Psi(a) > 0\})} \\ &= \overline{\{f(a) \in \mathfrak{R} \mid \Psi(a) > 0\}} \end{aligned} \quad (30)$$

As, $\{\theta \in \mathfrak{R} \mid (f(\Psi))(\theta) > 0\}$, so

$$\text{Supp}(f(\Psi)) = f(\text{Supp}(\Psi)) \quad (31)$$

Lemma 3. Suppose is a compact set, also as well as are continuous on, in addition

$$|f_1(a) - f_2(a)| < \zeta, \forall a \in G \quad (32)$$

Hence $|\sup_{a \in G_1} f_2(a)| < \zeta$ sup is valid for each compact set $G_1 \subset G$.

proof. As G_1 is a compact set also, f_1 as well as f_2 are continuous on G_1 , so there exist $a_0 \in G_1$, $\hat{a}_0 \in G_1$, in such a way that

$$\begin{aligned} f_1(a_0) &= \sup_{a \in G_1} f_1(a) \\ f_2(\hat{a}_0) &= \sup_{a \in G_1} f_2(a) \end{aligned} \quad (33)$$

Let $|f_1(a_0) - f_2(\hat{a}_0)| \geq \zeta$, so

$$\begin{aligned} f_1(a_0) - f_2(\hat{a}_0) &\leq -\zeta, \\ \text{or } f_1(a_0) - f_2(\hat{a}_0) &\geq \zeta \end{aligned} \quad (34)$$

In the first case of (34), since $f_1(\hat{a}_0) \leq f_1(a_0)$,

$$\begin{aligned} &f_1(\hat{a}_0) - f_2(\hat{a}_0) \\ &\leq f_1(a_0) - f_2(\hat{a}_0) \leq -\zeta \\ &\Rightarrow |f_1(\hat{a}_0) - f_2(\hat{a}_0)| \geq \zeta \end{aligned} \quad (35)$$

holds, that contradicts (32). In the second case (34), as $f_2(a_0) \leq f_2(\hat{a}_0)$, the following is obtained

$$\begin{aligned} &f_1(a_0) - f_2(a_0) \\ &\geq f_1(a_0) - f_2(\hat{a}_0) \geq \zeta \\ &\Rightarrow |f_1(a_0) - f_2(a_0)| \geq \zeta \end{aligned} \quad (36)$$

that contradicts (32). Hence, (34) is not correct, so $-\zeta < f_1(a_0) - f_2(\hat{a}_0) < \zeta$, hence $|f_1(a_0) - f_2(a_0)| < \zeta$, i.e. $|\sup_{a \in G_1} f_1(a) - \sup_{a \in G_1} f_2(a)| < \zeta$. The proof is finalized.

Theorem 1. Suppose $f : \mathfrak{R} \rightarrow \mathfrak{R}$ is a continuous function, so for each compact set $\varpi \subset \hat{Z}_0$ (the set of all the bounded Z-number set) also $\chi > 0$, there exists $a_\iota \in \hat{Z}_0$ $\iota = 1, 2, \dots, \gamma$, in such a manner that

$$d(f(\tilde{\vartheta}), \sum_{\iota=1}^{\gamma} f_{\iota}(\vartheta)a_{\iota}) < \chi \forall \vartheta \in \varpi, \forall \tilde{\vartheta} \in \mathfrak{R} \quad (37)$$

in which χ is a finite number.

proof. The subsequent results will lead to the proof of the Theorem.

Suppose $f : \mathfrak{R} \rightarrow \mathfrak{R}$, f can be extended by the extension principle to the Z-number function that is stated as $f : \hat{Z}_0 \rightarrow \hat{Z}$, also

$$f(\psi_1)(\omega_1) = \vee_{f(\vartheta)=\omega_1} \{\psi_1(\vartheta)\}, \quad \psi_1 \in \hat{Z}_0, \quad \omega_1 \in \mathfrak{R} \quad (38)$$

f is termed as the extended function. Furthermore, $cc(\mathfrak{R})$ is the set of bounded closed intervals of $\mathfrak{R}$. Clearly

$$\psi_1 \in \hat{Z}_0 \Rightarrow \forall \alpha \in (0, 1], [\psi_1]^\alpha \in cc(\mathfrak{R}) \quad (39)$$

Also,

$$\text{Supp}(\psi_1) \in cc(\mathfrak{R}) \quad (40)$$

Hence,

$$\text{Supp}(\psi_1) = [s_1(\psi_1), s_2(\psi_1)] \quad (41)$$

Theorem 2. If $f : \mathfrak{R} \rightarrow \mathfrak{R}$ is a continuous function, so for each compact set $\varpi \subset \hat{Z}_0$, $\tau > 0$ also arbitrary $\varepsilon > 0$, there exists $a_\iota \in \hat{Z}_0$, $\iota = 1, 2, \dots, \gamma$, in such a way that

$$d(f(\vartheta), \sum_{\iota=1}^{\gamma} f_{\iota}(\vartheta)a_{\iota}) < \tau, \quad \forall \vartheta \in \varpi \quad (42)$$

in which, τ is termed as a finite number. The lower as well as the upper limits of the α -level set of Z-number function approach to τ , however the center approaches to ε .

proof. Since $\varpi \subset \hat{Z}_0$ is a compact set, so by implementing Lemma 1, $\Psi \subset \mathfrak{R}$ is the compact set corresponding to ϖ . Furthermore, $\forall \varepsilon > 0$, also using the results of (Cybenko, 1989), there exists $a_\iota \in \mathfrak{R}$, $\iota = 1, 2, \dots, \gamma$, in such a way that

$$|f(\vartheta) - \sum_{i=1}^{\gamma} f_i(\vartheta)a_i| < \varepsilon, \quad \forall \vartheta \in \Psi \quad (43)$$

If $\tilde{f}(\vartheta) = \sum_{i=1}^{\gamma} f_i(\vartheta)a_i$, where $\vartheta \in \mathfrak{R}$, hence

$$|f(\vartheta) - \tilde{f}(\vartheta)| < \varepsilon, \quad \forall \vartheta \in \Psi \quad (44)$$

Theorem 3 will lead to the validation of (42).

Theorem 3. Suppose $\varpi \subset \widehat{Z}_0$ is compact, also Ψ is the corresponding compact set of ϖ and $f, f^E: \mathfrak{R} \rightarrow \mathfrak{R}$ are the continuous functions that,

$$|f(\vartheta) - f^E(\vartheta)| < \zeta, \quad \forall \vartheta \in \Psi \quad (45)$$

where $\zeta > 0$. So, $\forall \psi_1 \in \varpi$, $d(f(\psi_1) - f^E(\psi_1)) \leq \zeta$.

proof. Assume $\psi_1 \in \widehat{Z}$ also $\alpha \in (0, 1]$. Since f as well as f^E are continuous, then $[f(\psi_1)]^\alpha = f([\psi_1]^\alpha)$, and $[f^E(\psi_1)]^\alpha = f^E([\psi_1]^\alpha)$ are valid by Lemma 2. Hence, the following result is obtained using,

$$\begin{aligned} D_H([f(\psi_1)] - [f^E(\psi_1)]) &= D_H(f([\psi_1]^\alpha) - f^E([\psi_1]^\alpha)) \\ &= \sup_{|n|=1} \{ |s(n, f([\psi_1]^\alpha)) - s(n, f^E([\psi_1]^\alpha))| \} \end{aligned} \quad (46)$$

Since $n \in \mathfrak{R} : |n|=1$, the following is valid,

$$\begin{aligned} |s(n, f([\psi_1]^\alpha)) - s(n, f^E([\psi_1]^\alpha))| &= \sup \{ |nj| : j \in f([\psi_1]^\alpha) \} \\ &\quad - \sup \{ |nj| : j \in f^E([\psi_1]^\alpha) \} = \sup \{ |nf(\vartheta)| : \vartheta \in [\psi_1]^\alpha \} \\ &\quad - \sup \{ |nf^E(\vartheta)| : \vartheta \in [\psi_1]^\alpha \} \end{aligned} \quad (47)$$

Also, taking into consideration the conditions in the theorem, the following is obtained

$$|nf(\vartheta) - nf^E(\vartheta)| = |f(\vartheta) - f^E(\vartheta)| < \zeta, \quad \forall \vartheta \in [\psi_1]^\alpha \quad (48)$$

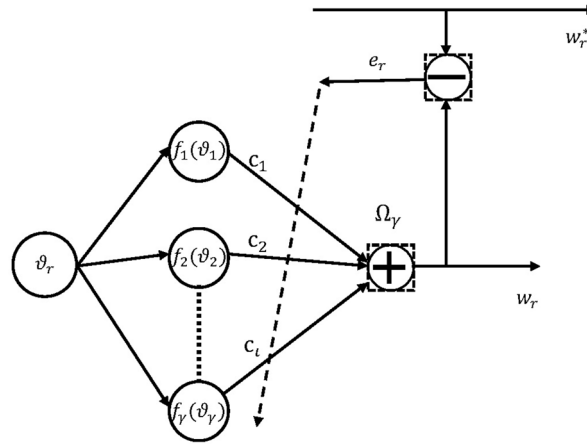
Hence, using (46), (47) as well as Lemma 3, the following relation is extracted which proves the theorem,

$$\begin{aligned} D_H([f(\psi_1)]^\alpha, [f^E(\psi_1)]^\alpha) &< \zeta \Rightarrow d(f(\psi_1), f^E(\psi_1)) \\ &= \sup_{\alpha \in (0, 1]} \{ D_H([f(\psi_1)]^\alpha, [f^E(\psi_1)]^\alpha) \} \leq \zeta \end{aligned} \quad (49)$$

Neural Network Approach for Z-number Parameter Approximation

In this section, a neural network is designed in order to demonstrate the fuzzy equation, see Figure 1. The input of the neural network is the Z-numbers ϑ_r , also the output is the Z-number w_r . The proposed neural network finds a_i in such a way that the output of the neural network w_r approaches to the desired output w_r^* .

Figure 1. Neural network model



The Z-number input ϑ_r is initially implemented to α -level as in (11)

$$[\vartheta_r]^\alpha = (\underline{\vartheta}_r^\alpha, \bar{\vartheta}_r^\alpha) \quad (50)$$

Afterwards, we have the following relation in the hidden unit

$$[\Omega_i]^\alpha = (f_i(\underline{\vartheta}_r)^\alpha, f_i(\bar{\vartheta}_r)^\alpha) \quad i = 1 \dots \gamma \quad (51)$$

The output of the neural network is

$$[w_r]^\alpha = \{ \sum_{i \in A} \Omega_i^\alpha \underline{a}_i^\alpha + \sum_{i \in B} \bar{\Omega}_i^\alpha \underline{a}_i^\alpha, \sum_{i \in A'} \bar{\Omega}_i^\alpha \bar{a}_i^\alpha + \sum_{i \in B'} \Omega_i^\alpha \bar{a}_i^\alpha \} \quad (52)$$

where $A = \{i | \underline{a}_i^\alpha \geq 0\}$, $B = \{i | \underline{a}_i^\alpha < 0\}$, $A' = \{i | \bar{a}_i^\alpha \geq 0\}$, $B' = \{i | \bar{a}_i^\alpha < 0\}$.

For training the weights, a cost function is defined for the Z-numbers. The training error is defined as follows

$$e_r = w_r^* - w_r \quad (53)$$

where $[w_r^*]^\alpha = (\underline{w}_r^{*\alpha}, \bar{w}_r^{*\alpha})$, $[w_r]^\alpha = (\underline{w}_r^\alpha, \bar{w}_r^\alpha)$, $[e_r]^\alpha = (\underline{e}_r^\alpha, \bar{e}_r^\alpha)$. The cost function is defined as

$$\begin{aligned}
\Phi_r &= \underline{\Phi}^\alpha + \overline{\Phi}^\alpha \\
\underline{\Phi}^\alpha &= \frac{1}{2} \left(\underline{w}_r^{*\alpha} - \underline{w}_r^\alpha \right)^2 \\
\overline{\Phi}^\alpha &= \frac{1}{2} \left(\overline{w}_r^{*\alpha} - \overline{w}_r^\alpha \right)^2
\end{aligned} \tag{54}$$

$\Phi_r \rightarrow 0$ signifies that $[w_r]^\alpha \rightarrow [w_r^*]^\alpha$.

The gradient technique is utilized in order to train the Z-number weight $a_i = (\underline{a}_i, \overline{a}_i)$. $\frac{\partial \Phi_r}{\partial a_i}$ is computed as follows

$$\begin{aligned}
\frac{\partial \Phi_r}{\partial a_i} &= \frac{\partial \Phi_r^\alpha}{\partial \underline{w}_r^\alpha} \frac{\partial \underline{w}_r^\alpha}{\partial \underline{a}_i} \frac{\partial \underline{a}_i^\alpha}{\partial a_i} + \frac{\partial \Phi_r^\alpha}{\partial \overline{w}_r^\alpha} \frac{\partial \overline{w}_r^\alpha}{\partial \overline{a}_i} \frac{\partial \overline{a}_i^\alpha}{\partial a_i} \\
&= -(\underline{w}_r^{*\alpha} - \underline{w}_r^\alpha) \left(\sum_{i \in A} \underline{\Omega}_i^\alpha + \sum_{i \in B} \overline{\Omega}_i^\alpha \right) \Gamma_1 \\
&\quad - (\overline{w}_r^{*\alpha} - \overline{w}_r^\alpha) \left(\sum_{i \in A} \overline{\Omega}_i^\alpha + \sum_{i \in B} \underline{\Omega}_i^\alpha \right) \Gamma_2
\end{aligned} \tag{55}$$

where $\Gamma_1 = (1 - \alpha, \underline{p}^\alpha)$ and $\Gamma_2 = (1 - \alpha, \overline{p}^\alpha)$.

The Z-number coefficient a_i is updated as

$$\begin{aligned}
\underline{a}_i(r+1) &= \underline{a}_i(r) - \eta \frac{\partial \Phi_r}{\partial \underline{a}_i} \\
\overline{a}_i(r+1) &= \overline{a}_i(r) - \eta \frac{\partial \Phi_r}{\partial \overline{a}_i}
\end{aligned} \tag{56}$$

where η is the training rate $\eta > 0$. In order to increase the training process, a momentum term is added as

$$\begin{aligned}
\underline{a}_i(r+1) &= \underline{a}_i(r) - \eta \frac{\partial \Phi_r}{\partial \underline{a}_i} + \gamma [\underline{a}_i(r) - \underline{a}_i(r-1)] \\
\overline{a}_i(r+1) &= \overline{a}_i(r) - \eta \frac{\partial \Phi_r}{\partial \overline{a}_i} + \gamma [\overline{a}_i(r) - \overline{a}_i(r-1)]
\end{aligned} \tag{57}$$

where $\gamma > 0$.

Learning Algorithm

- Step 1: Select the training rates $\eta > 0$, $\gamma > 0$, also the stopping criterion $\overline{\Phi} > 0$. The initial Z-number vector $A = (a_1, \dots, a_l)$ is chosen arbitrarily. The initial learning iteration is and the initial learning error is $\Phi = 0$.
- The following steps should be repeated for $\alpha = \alpha_1, \dots, \alpha_m$, till all the training data are implemented
 - Forward computation: Compute the α -level of the Z-number output w_r using the α -level of the Z-number input ϑ_r , and the Z-number connection weight A .
 - Back-propagation: Adjust Z-number parameters a_i , $i = 1, \dots, \gamma$, utilizing the cost function for the α -level of the Z-number output w_r , and the Z-number target output w_r^* .
 - Stopping criterion: Compute the cycle error Φ_r , $\Phi = \Phi + \Phi_r$, $r = r + 1$. In a case that $\Phi > \overline{\Phi}$, let $\Phi = 0$, a new training cycle is begun. Go to (a).

Numerical Example

In this section, some examples with applications are given to show how to apply fuzzy equations to model uncertain nonlinear systems.

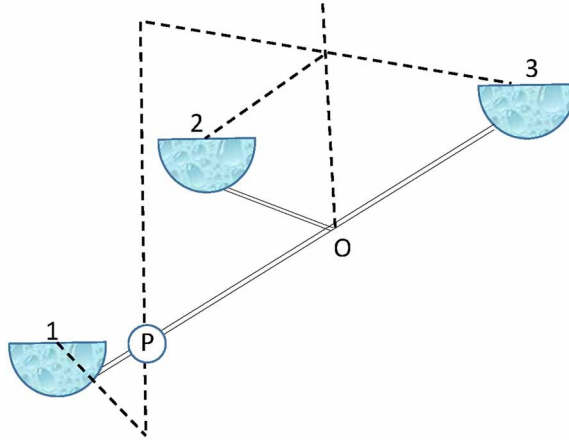
Example 1. Three tanks are linked to a pipeline at a constant O , see Figure 2. The water is pumped from one tank to the other two tanks. The system is defined as

$$O = C_1 G_1 \oplus C_2 G_2 \oplus C_3 G_3 \quad (58)$$

where $G_1 = \sqrt{39}$, $G_2 = \frac{1}{2}\sqrt{9}$, $G_3 = 9^2$ are taken to be the quantity of flow, ϑ is taken to be the elapsed time, O is considered to be the high of the pipe, C_1 , C_2 and C_3 are the characteristic coefficients of the pump, which are satisfied in the triangular Z-number uncertainty

$$\begin{aligned} C_1 &= [(3, 5, 8), p(0.81, 0.92, 1)], \\ C_2 &= [(2, 4, 7), p(0.72, 0.85, 0.9)], \\ C_3 &= [(2, 6, 8), p(0.82, 0.86, 0.91)] \end{aligned} \quad (59)$$

Figure 2. Tanks model



Three types of inputs are utilized for training the neural network, where weights are considered as the Z-number parameters of (58). The input data are given as

$$\vartheta = \{[(1, 5, 7), p(0.7, 0.8, 0.9)], [(10, 15, 19), p(0.7, 0.8, 0.91)], [(2, 3, 4), p(0.8, 0.91, 1)] \}$$

The input, ϑ is implemented to (58), and the corresponding outputs data are obtained as

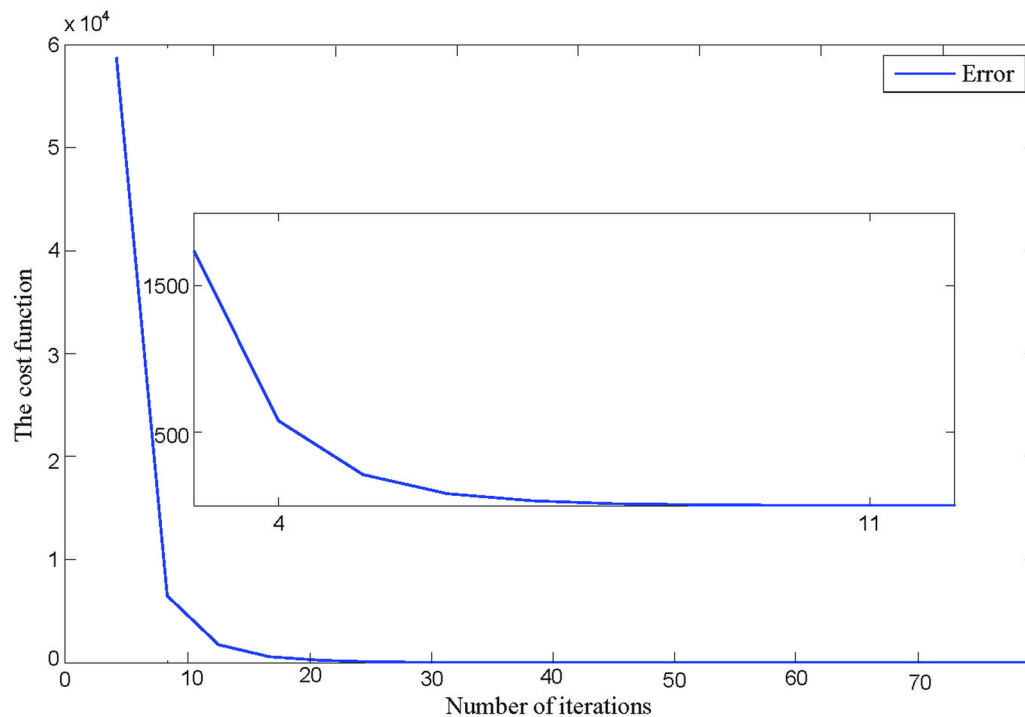
$$O = \left\{ \begin{aligned} &[(628.51, 668.91, 701.21), p(0.7, 0.8, 0.91)], \\ &[(3801.56, 4033.99, 4482.23), p(0.7, 0.82, 0.9)], \\ &[(2481.77, 2893.95, 3301.11), p(0.8, 0.86, 0.9)] \end{aligned} \right\}$$

These input-output pairs have been utilized for training the neural network. The weights are C_1 , C_2 , and C_3 . Table 1 demonstrates the training outcomes. It can be seen that the Z-number coefficients approach to their actual quantity. In this Table, r is taken to be the number of iterations. The approximation errors are shown in Figure 3.

Table 1. Estimation of Z-number coefficients

r		r		r	
1	[(6.27,9.17,11.19),p(0.6,0.8,0.85)]	1	[(5.36,8.28,12.35),p(0.71,0.8,0.92)]	1	[(7.43,10.49,12.37),p(0.7,0.8,0.92)]
2	[(5.91,8.75,10.64),p(0.72,0.8,0.87)]	2	[(4.89,7.77,11.85),p(0.8,0.86,0.94)]	2	[(6.82,9.91,11.87),p(0.6,0.81,0.87)]
	$\vdots$		$\vdots$		$\vdots$
70	[(3.02,5.01,8.03),p(0.81,0.92,1)]	70	[(2.06,4.05,7.03),p(0.72,0.85,0.9)]	70	[(2.04,6.06,8.05),p(0.82,0.86,0.91)]

Figure 3. Approximation errors



In order to change the Z-number into the fuzzy number, the following formula is utilized,

$$v = \frac{\int C \pi_p(C) dC}{\int \pi_p(C) dC} \quad (60)$$

Let $Z = [(6.27, 9.17, 11.19), p(0.6, 0.8, 0.85)]$, so $Z^v = (6.27, 9.17, 11.19; 0.7)$ therefore, $Z' = (\sqrt{0.7} \cdot 6.27, \sqrt{0.7} \cdot 9.17, \sqrt{0.7} \cdot 11.19)$. The results of Table 1 based of fuzzy numbers are demonstrated in Table 2.

Table 2. Estimation of fuzzy number coefficients

r		r		r	
1	(5.24,7.67,9.36)	1	(4.48,6.92,10.33)	1	(6.21,8.77,10.34)
2	(4.94,7.32,8.91)	2	(4.09,6.51,9.91)	2	(5.71,8.29,9.93)
70	(2.69,5.19,6.88)	70	(1.88,4.38,7.71)	70	(2.54,6.17,7.73)

The Z-number $Z=(c,p)=[(6.27,9.17,11.19),p(0.6,0.8,0.85)]$ has an advantage to fuzzy number (5.24,7.67,9.36) . It has more and various information, also the solution generated by Z-number is more precise.

Example 2. The heat source is placed in the left of the insulating materials, see Figure 4. The conductivity coefficient of these materials are $\frac{1}{C_M} = e^9$, $\frac{1}{C_N} = \sqrt{39}$, $\frac{1}{C_Q} = 9$, and 9 is taken to be the elapsed time. The system is defined as (Holman, 1997):

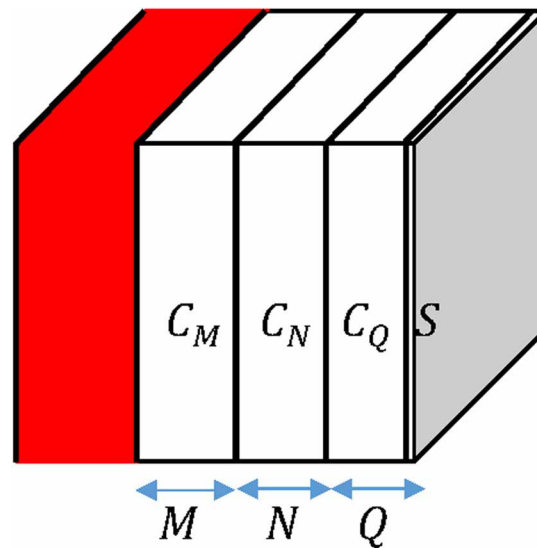
$$S = \frac{M}{C_M} \oplus \frac{N}{C_N} \oplus \frac{Q}{C_Q} \quad (61)$$

The real parameters are

$$\begin{aligned} M &= [(1, 5, 7), p(0.71, 0.8, 0.9)], \\ N &= [(7, 12, 15), p(0.82, 0.92, 1)], \\ Q &= [(2, 4, 8), p(0.81, 0.86, 0.9)] \end{aligned} \quad (62)$$

Three types of inputs are utilized for training the neural network. The input data are given as

Figure 4. Heat source



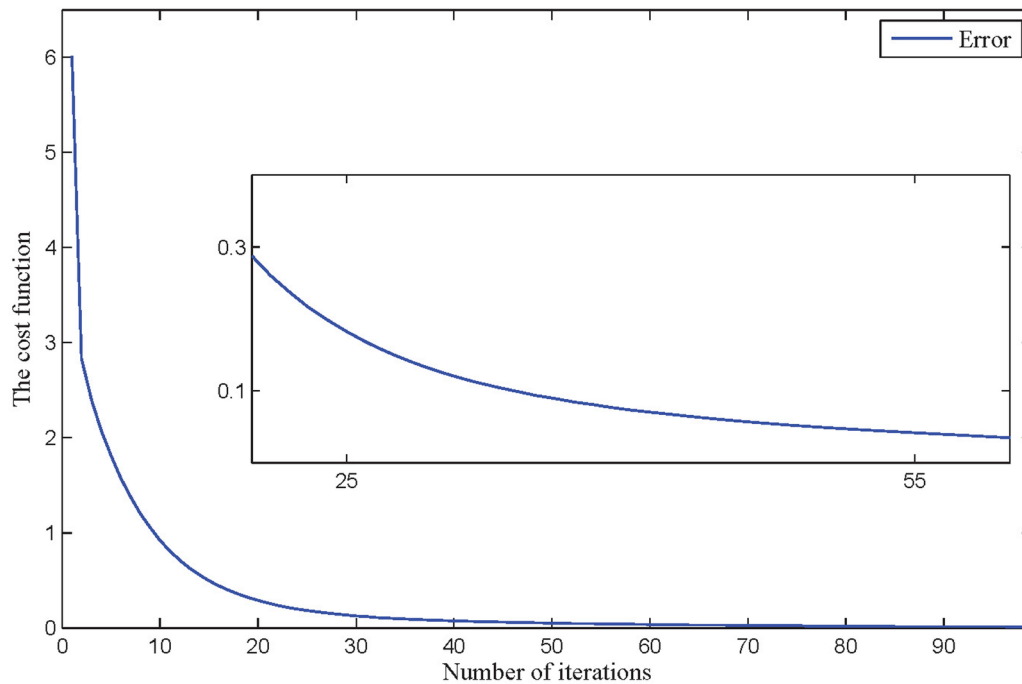
$$\mathfrak{S} = \{[(2, 4, 9), p(0.8, 0.9, 1)], [(8, 11, 13), p(0.7, 0.81, 0.9)], [(3, 4, 8), p(0.7, 0.8, 0.9)] \}$$

The input, $\mathfrak{S}$ is implemented to (61), and the corresponding outputs data are obtained as

$$S = \left\{ \begin{array}{l} [(430.61, 668.91, 701.21), p(0.7, 0.8, 0.9)], \\ [(1338.22, 1819.12, 2482.23), p(0.7, 0.81, 0.9)], \\ [(765.23, 964.15, 1039.77), p(0.8, 0.86, 0.9)] \end{array} \right\}$$

These input-output pairs have been utilized for training the neural network. The weights are M, N, and Q. The approximation errors are shown in Figure 5. Table 3 demonstrates the training outcomes. It can be seen that the Z-number coefficients approach to their actual quantity. In this Table, r is taken to be the number of iterations.

Figure 5. Approximation errors



The results of Table 3 based of fuzzy numbers are demonstrated in Table 4.

Table 3. Estimation of Z-number coefficients

r	M	r	N	r	Q
1	(5.38, 8.32, 11.23), p(0.81, 0.91, 1)]	1	[(10.23, 15.38, 18.41), p(0.7, 0.8, 0.9)]	1	[(6.31, 8.23, 12.42), p(0.7, 0.8, 0.92)]
2	[(4.13, 7.61, 10.49), p(0.7, 0.8, 0.9)]	2	[(9.71, 14.63, 17.65), p(0.8, 0.9, 1)]	2	[(5.61, 7.45, 11.44), p(0.6, 0.81, 0.87)]
	$\vdots$		$\vdots$		$\vdots$
97	[(1.02, 5.01, 7.03), p(0.6, 0.7, 0.8)]	97	[(7.04, 12.07, 15.05), p(0.7, 0.8, 0.9)]	97	[(2.06, 4.05, 8.04), p(0.82, 0.86, 0.91)]

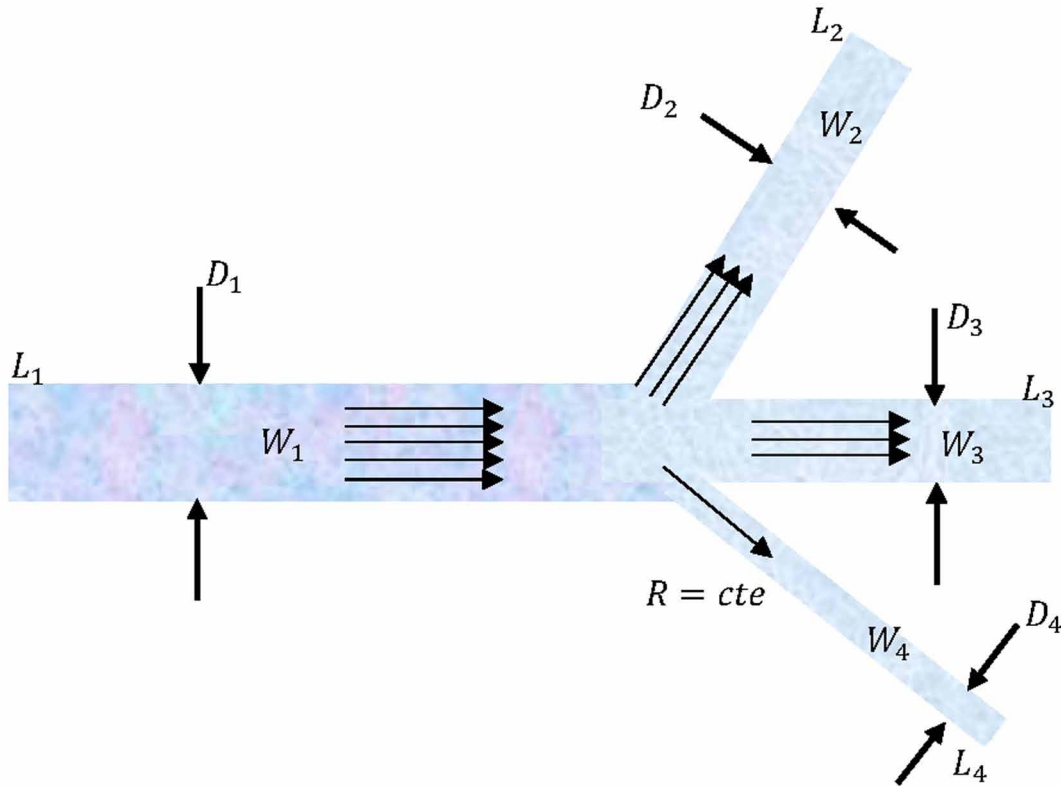
Table 4. Estimation of fuzzy number coefficients

r	M	N	Q
1	(4.04, 6.97, 9.99)	(9.08, 13.98, 17.31)	(4.99, 7.17, 11.06)
2	(3.11, 5.93, 9.51)	(7.99, 13.51, 16.01)	(4.12, 6.12, 10.11)
97	(1.59, 4.12, 6.51)	(6.28, 11.31, 14.75)	(2.55, 3.18, 7.65)

Example 3. The water in the pipe L_1 divides into three pipes L_2 , L_3 and L_4 , see Figure 6. The areas of the pipes are z-numbers,

$$\begin{aligned} D_2 &= [(2, 5, 8), p(0.7, 0.8, 0.9)], \\ D_3 &= [(7, 8, 9), p(0.8, 0.92, 1)], \\ D_4 &= [(3, 6, 9), p(0.8, 0.86, 0.9)] \end{aligned} \quad (63)$$

Figure 6. Water in the pipe model



The water velocities in the pipes are controlled by the valves parameter ϑ , $W_2 = e^{\vartheta}$, $W_3 = \vartheta$ and $W_4 = \sqrt{\vartheta}$. The system is defined as (Streeter, 1999),

$$R = D_2 W_2 \oplus D_3 W_3 \oplus D_4 W_4 \quad (64)$$

The input data are given as

$$\mathcal{G} = \{[(1, 5, 9), p(0.7, 0.8, 0.9)], [(8, 10, 12), p(0.7, 0.81, 0.9)], [(2, 4, 7), p(0.8, 0.9, 1)]\}$$

The corresponding outputs data are obtained as

$$R = \left\{ \begin{array}{l} [(532.74, 741.99, 821.13), p(0.8, 0.9, 1)], \\ [(1465.11, 1902.31, 2552.31), p(0.7, 0.8, 0.9)], \\ [(841.31, 1011.32, 1187.62), p(0.8, 0.85, 0.9)] \end{array} \right\}$$

These input-output pairs have been utilized for training the neural network. The weights are D_2 , D_3 , and D_4 . The approximation errors are shown in Figure 7. Table 5 demonstrates the training outcomes. It can be seen that the Z-number coefficients approach to their actual quantity. In this Table, r is taken to be the number of iterations.

Figure 7. Approximation errors

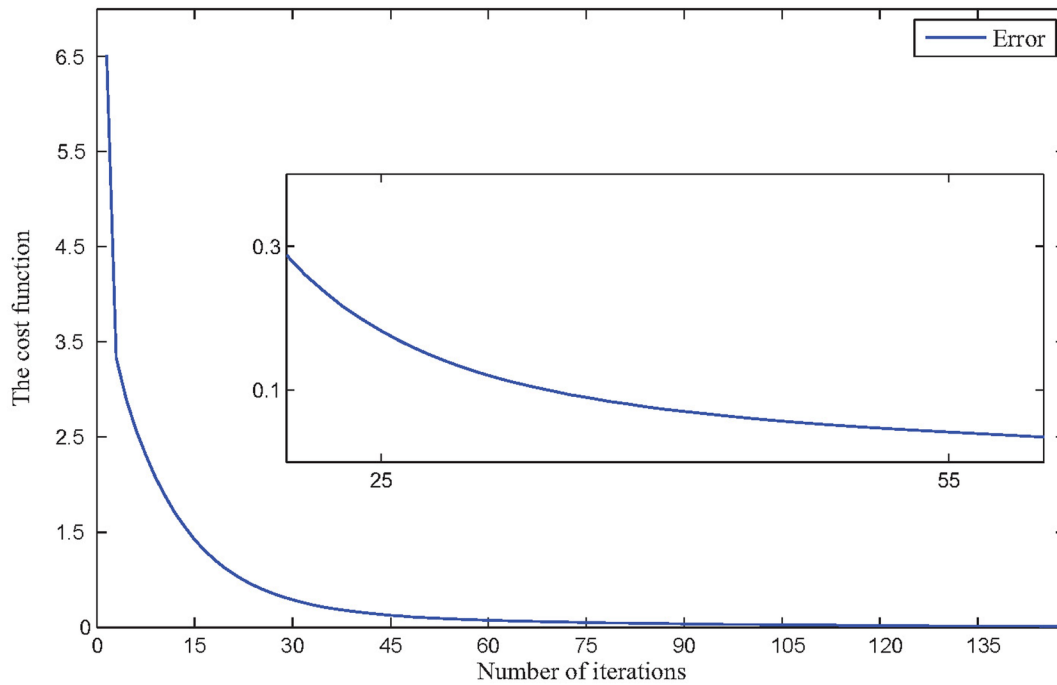


Table 5. Estimation of Z-number coefficients

r	D_2	r	D_3	r	D_4
1	[(4.45, 8.48, 12.44), p(0.82, 0.92, 1)]	1	[(11.33, 13.42, 15.41), p(0.7, 0.8, 0.9)]	1	[(5.27, 7.39, 10.41), p(0.7, 0.82, 0.92)]
2	[(3.78, 7.68, 11.73), p(0.6, 0.7, 0.85)]	2	[(10.82, 12.91, 14.81), p(0.8, 0.9, 1)]	2	[(4.65, 6.79, 9.81), p(0.6, 0.8, 0.87)]
	$\vdots$		$\vdots$		$\vdots$
169	[(2.02, 5.01, 8.03), p(0.71, 0.8, 0.96)]	169	[(7.05, 8.04, 9.03), p(0.7, 0.8, 0.9)]	169	[(3.05, 6.05, 9.03), p(0.81, 0.86, 0.91)]

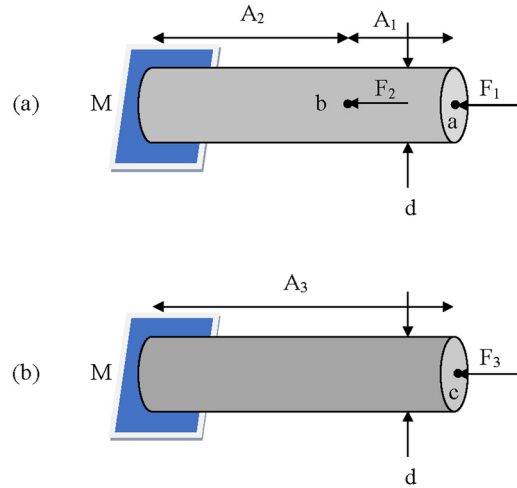
Example 4. The deformation of a solid cylindrical rod depends to the stiffness S , forces on it F_1, F_2, F_3 , the positions of the forces A_1, A_2, A_3 and the diameter of the rod $R = \frac{\pi}{4}d^2$. (Tahavvor & Yaghoubi, 2012) see Figure 8. The area of the rod is $R = \frac{\pi}{4}d^2$. The external forces are the function of ϑ , $F_1 = \vartheta$, $F_2 = \sqrt{29}$, $F_3 = 3\vartheta$ (Beer, 1992). The system is defined as (Beer & Johnston, 1992),

$$\frac{-A_1(F_1 + F_2 + F_3)}{R * SS} \oplus \frac{-A_2(F_2 + F_3)}{R * SS} \oplus \frac{-A_3 F_3}{R * SS} = Y \quad (65)$$

where $d=0.02$, $SS=2 \times 10^3$, and,

$$\begin{aligned} \frac{-A_1}{R * SS} &= [(3, 5, 9), p(0.7, 0.81, 0.91)], \\ \frac{-A_2}{R * SS} &= [(6, 8, 10), p(0.8, 0.9, 1)], \\ \frac{-A_3}{R * SS} &= [(3, 7, 11), p(0.8, 0.85, 0.9)] \end{aligned} \quad (66)$$

Figure 8. Solid cylindrical rod



The input data are given as

$$\vartheta = \{[(5, 7, 9), p(0.8, 0.9, 1)], [(9, 11, 13), p(0.7, 0.8, 0.9)], [(1, 4, 9), p(0.7, 0.8, 0.9)]\}$$

The corresponding outputs data are obtained as

$$R = \left\{ \begin{aligned} &[(678.91, 904.34, 1102.11), p(0.8, 0.9, 1)], \\ &[(1994.85, 2201.11, 2487.88), p(0.7, 0.8, 0.9)], \\ &[(452.82, 693.41, 891.12), p(0.8, 0.85, 0.9)] \end{aligned} \right\}$$

These input-output pairs have been utilized for training the neural network. The weights are $\frac{-A_1}{R * SS}$, $\frac{-A_2}{R * SS}$, and $\frac{-A_3}{R * SS}$. For comparing our results, the fuzzy cubic spline technique is utilized (Floreno & Novelli, 1996). Both techniques, the neural network method (the approach suggested in this paper) and the fuzzy cubic spline method are effective in approximating the Z-number coefficients of fuzzy equations. However, the estimated errors related to the neural network technique are minor than the fuzzy cubic spline technique, see Figure 9. Table 6 demonstrates the training outcomes. It can be seen that the Z-number coefficients approach to their actual quantity. In this Table, is taken to be the number of iterations.

Figure 9. Approximation errors of neural network technique and fuzzy cubic spline technique

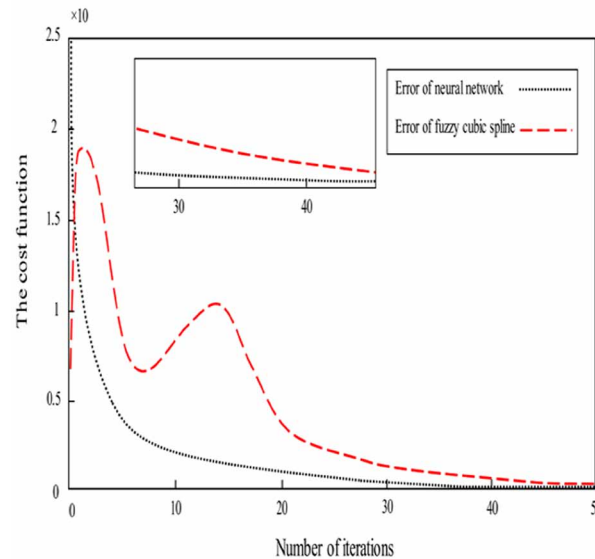


Table 6. Estimation of z – number coefficients

r	$\frac{-A_1}{R * SS}$	r	$\frac{-A_2}{R * SS}$	r	$\frac{-A_3}{R * SS}$
1	[(7.38, 9.42, 13.41), p(0.8, 0.9, 1)]	1	[(9.41, 12.39, 13.35), p(0.8, 0.9, 1)]	1	[(6.38, 10.41, 14.36), p(0.7, 0.82, 0.9)]
2	[(6.69, 8.72, 12.72), p(0.6, 0.7, 0.8)]	2	[(8.78, 11.81, 12.69), p(0.8, 0.9, 1)]	2	[(5.72, 9.69, 13.79), p(0.6, 0.8, 0.85)]
					$\vdots$
169	[(3.03, 5.02, 9.03), p(0.7, 0.8, 0.9)]	169	[(6.03, 8.05, 10.04), p(0.7, 0.8, 0.9)]	169	[(3.04, 7.05, 11.06), p(0.8, 0.86, 0.9)]

FUTURE RESEARCH DIRECTIONS

The investigation of the stability of the proposed training algorithm will be the further work.

CONCLUSION

In this book chapter, the neural network approach is utilized in order to estimate the Z-number coefficients of fuzzy equations. The structure of the neural network is based on fuzzy equations. The backpropagation technique is implemented for training the neural network. Simulation results reveal the efficiency of the suggested technique.

REFERENCES

- Abbasbandy, S. (2006). The application of homotopy analysis method to nonlinear equations arising in heat transfer. *Physics Letters. [Part A]*, 360(1), 109–113.
- Abbasbandy, S., & Ezzati, R. (2006). Newton's method for solving a system of fuzzy nonlinear equations. *Applied Mathematics and Computation*, 175(1), 1189–1199.
- Aliev, R. A., Alizadeh, A. V., & Huseynov, O. H. (2015). The arithmetic of discrete Z-numbers. *Inform. Sci.*, 290, 134–155.
- Aliev, R. A., Pedrycz, W., Kreinovich, V., & Huseynov, O. H. (2016). The general theory of decisions. *Inform. Sci.*, 327(1), 125–148.
- Allahviranloo, T., Otadi, M., & Mosleh, M. (2007). Iterative method for fuzzy equations. *Soft Computing*, 12, 935–939.
- Barthelmann, V., Novak, E., & Ritter, K. (2000). High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics*, 12, 273–288.
- Bede, B., & Stefanini, L. (2013). Generalized differentiability of fuzzy-valued functions. *Fuzzy Sets and Systems*, 230, 119–141.
- Beer, F. P., & Johnston, E. R. (1992). *Mechanics of materials* (2nd ed.). McGraw-Hill.
- Boffi, D., & Gastaldi, L. (2006). Interpolation estimates for edge finite elements and application to band gap computation. *Applied Numerical Mathematics*, 56, 1283–1292.
- Buckley, J., & Eslami, E. (1997). Neural net solutions to fuzzy problems: The quadratic equation. *Fuzzy Sets and Systems*, 86, 289–298.
- Buckley, J., & Qu, Y. (1990). Solving linear and quadratic fuzzy equations. *Fuzzy Sets and Systems*, 35, 43–59.
- Cybenko, G. (1989). Approximation by Superposition of Sigmoidal Activation Function. *Mathematics of Control, Signals, and Systems*, 2, 303–314.
- Floreno, M. C., & Novelli, G. (1996). Implementing fuzzy polynomial interpolation (FPI) and fuzzy linear regression (LFR). *Le Matematiche*, 51, 59–76.
- Friedman, M., Ming, M., & Kandel, A. (1998). Fuzzy linear systems. *Fuzzy Sets and Systems*, 96, 201–209.
- Gardashova, L. A. (2014). Application of operational approaches to solving decision making problem using Z-Numbers. *Journal of Applied Mathematics*, 5, 1323–1334.

Goetschel, R., & Voxman, W. (1986). Elementary calculus. *Fuzzy Sets and Systems*, 18, 31–43.

Holman, J. P. (1997). *Heat transfer*. McGraw-Hill.

Ito, Y. (2001). Independence of unscaled basis functions and finite mappings by neural networks. *Math. Sci.*, 26, 117–126.

Jafari, R., & Razvarz, S. (2017). Solution of Fuzzy Differential Equations using Fuzzy Sumudu Transforms. *IEEE International Conference on Innovations in Intelligent Systems and Applications*, 84-89.

Jafari, R., & Razvarz, S. (2018). Solution of fuzzy differential equations using fuzzy sumudu transforms. *Mathematical and Computational Applications*, 1-15.

Jafari, R., Razvarz, S., & Gegov, A. (2018a). A new computational method for solving fully fuzzy nonlinear systems. *Computational Collective Intelligence: 10th International Conference, ICCCI 2018*, 503-512.

Jafari, R., Razvarz, S., & Gegov, A. (2019a). Neural Network Approach to Solving Fuzzy Nonlinear Equations using Z-Numbers. *IEEE Transactions on Fuzzy Systems*. Advance online publication. doi:10.1109/TFUZZ.2019.2940919

Jafari, R., Razvarz, S., & Gegov, A. (2019b). Solving Differential Equations with Z-Numbers by Utilizing Fuzzy Sumudu Transform. *Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing*, Springer, 1125-1138.

Jafari, R., Razvarz, S., Gegov, A., & Paul, S. (2018b). Fuzzy modeling for uncertain nonlinear systems using fuzzy equations and Z-numbers. *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence*, 66-107.

Jafari, R., & Yu, W. (2015a). Artificial neural network approach for solving strongly degenerate parabolic and burgers-fisher equations. *12th International Conference on Electrical Engineering, Computing Science and Automatic Control*. doi:10.1109/ICEEE.2015.7357914

Jafari, R., & Yu, W. (2015b). Fuzzy Control for Uncertainty Nonlinear Systems with Dual Fuzzy Equations. *Journal of Intelligent & Fuzzy Systems*, 29, 1229–1240.

Jafari, R., & Yu, W. (2015c). Uncertainty Nonlinear Systems Control with Fuzzy Equations. *IEEE International Conference on Systems, Man, and Cybernetics*, 2885-2890.

Jafari, R., & Yu, W. (2015d). Uncertainty nonlinear systems modeling with fuzzy equations. *Proceedings of the 16th IEEE International Conference on Information Reuse and Integration (IRI '15)*, 182-188.

Jafari, R., & Yu, W. (2017a). Fuzzy Differential Equation for Nonlinear System Modeling with Bernstein Neural Networks. *IEEE Access: Practical Innovations, Open Solutions*. Advance online publication. doi:10.1109/ACCESS.2017.2647920

Jafari, R., & Yu, W. (2017b). Uncertain nonlinear system control with fuzzy differential equations and Z-numbers. *18th IEEE International Conference on Industrial Technology*, 890 -895, doi:10.1109/ICIT.2017.7915477

Jafari, R. Yu, W., & Li, X. (2016). Solving Fuzzy Differential Equation with Bernstein Neural Networks. *IEEE International Conference on Systems, Man, and Cybernetics*, 1245-1250.

Jafarian, A., Measoomy nia, S., & Jafari. R. (2012). Solving Fuzzy Equations Using Neural Nets with a New Learning Algorithm. *Journal of Advances in Computer Research*, 3, 33–45.

Jafarian, A., & Jafari, R. (2012). Approximate solutions of dual fuzzy polynomials by feed-back neural networks. *J. Soft Comput. Appl.*, 2012. Advance online publication. doi:10.5899/2012/jsca-00005

Jafarian, A., Jafari, R., Mohamed Al Qurashi, M., & Baleanu, D. (2016). A novel computational approach to approximate fuzzy interpolation polynomials. *SpringerPlus*, 5, 1428–1435. doi:10.1186/40064-016-3077-5 PMID:27625982

Kajani, M., Asady, B., & Vencheh, A. (2005). An iterative method for solving dual fuzzy nonlinear equations. *Applied Mathematics and Computation*, 167, 316–323.

Kajani, M., & Lupulescu, B. V. (2009). On a class of fuzzy functional differential equations. *Fuzzy Sets and Systems*, 160, 1547–1562.

Kang, B., Wei, D., Li, Y., & Deng, Y. (2012a). A method of converting Z-number to classical fuzzy number. *Journal of Information and Computational Science*, 9, 703–709.

Kang, B., Wei, D., Li, Y., & Deng, Y. (2012b). Decision making using Z—Numbers under uncertain environment. *Journal of Computer Information Systems*, 8, 2807–2814.

Llanas, B., & Sainz, F. J. (2006). Constructive approximate interpolation by neural networks. *Journal of Computational and Applied Mathematics*, 188, 283–308.

Mamdani, E. H. (1976). Application of fuzzy algorithms for control of simple dynamic plant. *IEE Proceedings. Control Theory and Applications*, 121(12), 1585–1588.

Mastylo, M. (2010). Interpolation estimates for entropy numbers with applications to non-convex bodies. *J. Approx. Theory*, 162, 10–23.

Mazandarani, M., & Kamyad, A. V. (2013). Modified fractional Euler method for solving fuzzy fractional initial value problem. *Communications in Nonlinear Science and Numerical Simulation*, 18, 12–21.

Mosleh, M. (2013). Evaluation of fully fuzzy matrix equations by fuzzy neural network. *Applied Mathematical Modelling*, 37, 6364–6376.

Neidinger, R. D. (2009). Multivariable interpolating polynomials in newton forms. Joint mathematics meetings, 5-8.

Olver, P. J. (2006). On multivariate interpolation. *Studies in Applied Mathematics*, 116, 201–240.

Pederson, S., & Sambandham, M. (2008). The Runge-Kutta method for hybrid fuzzy differential equation. *Nonlinear Analysis. Hybrid Systems*, 2, 626–634.

Razvarz, S., & Jafari, R. (2017a). ICA and ANN Modeling for Photocatalytic Removal of Pollution in Wastewater. *Mathematical and Computational Applications*, 22, 38–48.

Razvarz, S., & Jafari, R. (2017b). Intelligent Techniques for Photocatalytic Removal of Pollution in Wastewater. *Journal of Electrical Engineering*, 5, 321–328. doi:10.17265/2328- 2223/2017.06.004

Razvarz, S., Jafari, R., Gegov, A., Yu, W., & Paul, S. (2018). Neural network approach to solving fully fuzzy nonlinear systems. *Fuzzy modeling and control Methods Application and Research*, 45-68.

- Razvarz, S., Jafari, R., Yu, W., & Khalili, A. (2017). *PSO and NN Modeling for Photocatalytic Removal of Pollution in Wastewater*. 14th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico City, Mexico.
- Salahshour, S., Allahviranloo, T., & Abbasbandy, S. (2012). Solving fuzzy fractional differential equations by fuzzy Laplace transforms. *Communications in Nonlinear Science and Numerical Simulation*, 17, 1372–1381.
- Schroeder, H., Murthy, V. K., & Krishnamurthy, E. V. (1991). Systolic algorithm for polynomial interpolation and related problems. *Parallel Computing*, 493–503.
- Streeter, V. L., Wylie, E. B., & Benjamin, E. (1999). *Fluid mechanics*. Mc. Graw-Hill Book Company.
- Szabados, J., & Vertesi, P. (1990). *Interpolation of Functions*. World Scientific.
- Tahavvor, A., & Yaghoubi, M. (2012). Analysis of natural convection from a column of cold horizontal cylinders using artificial neural network. *Applied Mathematical Modelling*, 36, 3176–3188.
- Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15, 116–132.
- Tikhomirov, V. M. (1990). Approximation theory. In *Encyclopaedia of Mathematical Sciences*. Springer-Verlag.
- Wang, H., & Liu, Y. (2011). Existence results for fuzzy integral equations of fractional order. *Int. J. Math. Anal.*, 5, 811–818.
- Waziri, M., & Majid, Z. (2012). A new approach for solving dual fuzzy nonlinear equations using Broyden's and Newton's methods. *Advances in Fuzzy Systems*, 682087, 1–5.
- Xiu, D., & Hesthaven, J. S. (2005). High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27, 1118–11139.
- Yang, L. B., & Gao, Y. Y. (1993). *Fuzzy Mathematics-Theory and its Application*. South China University of Technology.
- Yu, W., & Jafari, R. (2019). *Fuzzy modeling and control with fuzzy equations and z-number*, *IEEE Press series on systems science and engineering*. Wiley-IEEE Press. John Wiley & Sons, Inc.
- Yu, W., & Li, X. (2004). Fuzzy identification using fuzzy neural networks with stable learning algorithms. *IEEE Transactions on Fuzzy Systems*, 12(3), 411–420.
- Zadeh, L. A. (2005). Toward a generalized theory of uncertainty (GTU) an outline. *Inform. Sci.*, 172, 1–40.
- Zadeh, L. A. (2006). Generalized theory of uncertainty (GTU)-principal concepts and ideas. *Computational Statistics & Data Analysis*, 51, 15–46.
- Zadeh, L. A. (2011). A note on Z-numbers. *Information Sciences*, 181, 2923–2932.
- Zolic, A. (2008). *Numerical mathematics*. Faculty of Mathematics: Belgrade.

KEY TERMS AND DEFINITIONS

Algorithm: An algorithm in mathematics is a procedure, a description of a set of steps that can be used to solve a mathematical computation.

Function: A relation between sets that associates to every element of a first set exactly one element of the second set.

Neural Network: A neural network is a set of layers (a layer has set of neurons) stacked together sequentially.

Nonlinear System: A system in which the change of the output is not proportional to the change of the input.

Polynomial: An expression of more than two algebraic terms, especially the sum of several terms that contain different powers of the same variable(s).

Z-Number: Z-number associated with an uncertain variable X as an ordered pair of fuzzy numbers (A, B) where A represents a value of the variable and B represents an idea of certainty.

Use of Grey Numbers for Evaluating a System's Performance Under Fuzzy Conditions

Michael Voskoglou

Graduate Technological Educational Institute of Western Greece, Greece

INTRODUCTION

In Management a **system** is understood to be a set of interacting components forming an integrated whole and working together for achieving a common target. A factory, a hospital or a bank, constitute standard examples of systems, as well as a physical or biological system, an abstract knowledge system, etc. As a multi- perspective domain systems' theory serves as a bridge for an interdisciplinary dialogue between autonomous areas of study (Balley, 1994, Backlund, 2000). The assessment of a system's performance constitutes a very important part of this theory, because it enables the designers of the system to correct its weaknesses and therefore to increase its effectiveness.

When the performance of a system's components is evaluated by numerical scores, then the traditional way for assessing the system's **mean performance** is the calculation of the average of those scores. However, in order to comfort the reviewer's existing uncertainty about the exact value of the numerical scores corresponding to each of the system's components, frequently in practice the assessment is made not by numerical scores but by linguistic grades, like excellent, very good, good, etc. This involves a degree of fuzziness and makes the calculation of the mean value of those grades impossible.

A popular in such cases method for evaluating the overall system's performance is the calculation of the **Grade Point Average (GPA) index** (e.g. see Voskoglou, 2017b, Chapter 6, p.125). However, GPA is a weighted average in which greater coefficients (weights) are assigned to the higher grades, which means that it reflects not the mean, as we wish, but the **quality performance** of the system.

Analysis of the objectives and of the scope of the present work: In order to overcome the above difficulties, we have utilized in earlier works the system's **total uncertainty** under fuzzy conditions (created by the qualitative assessment of its components) as a measure of its effectiveness (Voskoglou, 2017b, Chapter 5). This manipulation is based on a fundamental principle of the Information Theory according to which the reduction of a system's uncertainty is connected to the increase of information obtained by a system's activity. In other words, lower uncertainty indicates a greater amount of information and therefore a better system's performance with respect to the corresponding activity. However, this method needs laborious calculations, cannot give a precise qualitative characterization of a system's performance and, most importantly, it is applicable for comparing the performance of two different systems with respect to a common activity only under the assumption that the existing uncertainty is the same in the two systems before the activity.

For this reason, **Fuzzy Numbers (FNs)** have been also used in later works for assessing a system's mean performance under fuzzy conditions (Voskoglou, 2017a). However it was observed that, although the calculation of three components is needed for expressing the mean value of the qualitative grades in the form of a **Triangular FN (TFN)**, only the middle component is used for defuzzifying it. The above

DOI: 10.4018/978-1-7998-3479-3.ch023

observation suggests the search for a method analogous to the use of TFNs that possibly reduces the required computational burden. This search led us to utilizing **Grey Numbers (GNs)** (Liu & Lin, 2010) as an alternative tool for assessing a system's mean performance with qualitative grades (Voskoglou & Theodorou, 2017). Although this new method has been proved to be equivalent with the use of TFNs, it needs the calculation of two components only (instead of three) for expressing the mean value of the qualitative grades in the form of a GN.

Synthesis of the present work: Here we are going to present these two innovative methods and to prove their equivalence. The rest of the Chapter is formulated as follows: In the second (background) section the assessment method with the TFNs is sketched and the necessary information about GNs is given, needed for the understanding of the present work. The alternative assessment method using GNs is developed in the third section and its equivalence is proved to the method with TFNs. Examples illustrating those methods are presented in the fourth section and the perspectives of future research on the subject are discussed in the fifth section. Finally, the sixth section includes the conclusions of the present research.

BACKGROUND

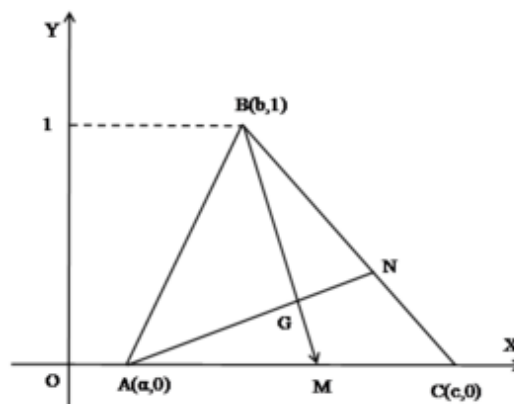
Triangular Fuzzy Numbers (TFNs)

The reader is supposed to be familiar with the basic principles of the theory of **Fuzzy Sets (FSs)** and the books of Klir & Folger (1988) and of Dubois & Prade (1980) are proposed as general references on the subject.

A **Fuzzy Number (FN)**, say A , is a **FS** on the set $\mathbf{R}$ of the real numbers, which is **normal** (i.e. there exists x in $\mathbf{R}$ such that $m_A(x) = 1$) and **convex** (i.e. all its **α -cuts** $A^\alpha = \{x \in U: m_A(x) \geq \alpha\}$, α in $[0, 1]$ are closed real intervals) and whose **membership function** $y = m_A(x)$ is a piecewise continuous function. For general facts on FNs the reader may study the classical on the subject book of Kaufmann and Gupta (1994)

A TFN (a, b, c) , with a, b, c real numbers such that $a < b < c$ is the simplest form of a FN representing mathematically the fuzzy statement that “the value of b lies in the interval $[a, c]$ ”. The membership function $y = m(x)$ of (a, b, c) is zero outside the interval $[a, c]$, while its graph in $[a, c]$ consists of two straight line segments forming a triangle with the OX axis (Figure 1).

Figure 1. Graph and Centre of Gravity of the TFN (a, b, c)



Therefore we have:

$$y = m(x) = \begin{cases} \frac{x-a}{b-a}, & x \in [a, b] \\ \frac{c-x}{c-b}, & x \in [b, c] \\ 0, & x < a \text{ or } x > c \end{cases}$$

Using elementary methods of Analytic Geometry it is straightforward to check (Voskoglou, 2017a, Proposition 8) that the coordinates (X, Y) of the *Centre of Gravity (COG)* of the graph of the TFN $A = (a, b, c)$, being the intersection point G of the medians of the corresponding triangle (Figure 1), are calculated by the formulas

$$X(A) = \frac{a+b+c}{3}, Y(A) = \frac{1}{3}(1).$$

The first of formulas (1) can be used to defuzzify the TFN A (*COG defuzzification technique*, e.g. see van Broekhoven, & De Baets, 2006), i.e. to represent it by a crisp number.

The two well known and equivalent to each other methods of defining *arithmetic operations* on FNs (Kaufmann & Gupta, 1991) lead to the following simple rules for the *addition* and *subtraction* of TFNs:

Let $A = (a, b, c)$ and $B = (a_p, b_p, c_p)$ be two TFNs. Then one defines:

- The **sum** $A + B = (a+a_p, b+b_p, c+c_p)$.
- The **difference** $A - B = A + (-B) = (a-c_p, b-b_p, c-a_p)$, where $-B = (-c_p, -b_p, -a_p)$ is defined to be the *opposite* of B .

On the contrary, the **product** and the **quotient** of A and B are FNs, which are not TFNs in general, apart from some special cases.

The following two **scalar operations** can be also defined:

- $k + A = (k+a, k+b, k+c), k \in \mathbf{R}$
- $kA = (ka, kb, kc)$, if $k > 0$ and $kA = (kc, kb, ka)$, if $k < 0$.

The Assessment Method with TFNs

Let $A_i, i = 1, 2, \dots, n$ be TFNs, where n is a non negative integer, $n \geq 2$. Then, the **mean value** of the A_i 's is defined to be the TFN $A = \frac{1}{n}(A_1 + A_2 + \dots + A_n)$.

The qualitative grades $A = \text{excellent}$, $B = \text{very good}$, $C = \text{good}$, $D = \text{fair}$ and $F = \text{unsatisfactory (failed)}$ are considered for the assessment of a system's performance and a scale of numerical scores from 1 – 100 is assigned to them as follows: $A (85 - 100)$, $B (75 - 84)$, $C (60 - 74)$, $D (50 - 59)$ and $F (0 - 49)$ ¹.

Then, each of the above grades can be represented by a TFN, denoted for simplicity by the same letter, in the following way: A = (85, 92.5, 100), B = (75, 79.5, 84), C (60, 67, 74), D (50, 54.5, 59) and F (0, 24.5, 49).

The middle entry of each of the above TFNs is equal to the average of its other two entries. In other words, if

$$A(a_1, b_1, c_1), B(a_2, b_2, c_2), \dots, F(a_5, b_5, c_5), \text{ then } b_i = \frac{a_i + c_i}{2}, i = 1, 2, 3, 4, 5.$$

In order to assess the total system's effectiveness, the performance of each of the system's components is evaluated by one of the above five qualitative grades, which means that one of the TFNs A, B, C, D, F is assigned to each component.

Let n be the total number of the system's components and let n_x denote the number of the components corresponding to the TFN X, where $X = A, B, C, D, F$. Then the mean value M of all those TFNs is equal to the TFN

$$M(a, b, c) = \frac{1}{n} (n_A A + n_B B + n_C C + n_D D + n_F F) \quad (2).$$

Since the calculation of the mean value of the qualitative grades is not possible, it looks logical to consider the TFN M as a fuzzy tool for evaluating the system's mean performance. Then, calculating by the first of formulas (1) the x-coordinate $X(M)$ of the COG of the graph of M, one obtains a crisp quantity for assessing this performance.

Remark: The mean value M is a linear combination of the TFNs A, B, C, D, F of the form $M = k_1 A + k_2 B + k_3 C + k_4 D + k_5 F$, with k_i non negative rational numbers, $i = 1, 2, 3, 4, 5$. Consequently, if $A(a_1, b_1, c_1), B(a_2, b_2, c_2), \dots, F(a_5, b_5, c_5)$ and $M(a, b, c)$, then $M(a, b, c)$

$$M = \sum_{i=1}^5 k_i (a_i, b_i, c_i) = \left(\sum_{i=1}^5 k_i a_i, \sum_{i=1}^5 k_i b_i, \sum_{i=1}^5 k_i c_i \right).$$

Therefore,

$$X(M) = \frac{\sum_{i=1}^5 k_i a_i + \sum_{i=1}^5 k_i b_i + \sum_{i=1}^5 k_i c_i}{3} = \sum_{i=1}^5 k_i \frac{a_i + b_i + c_i}{3} = \sum_{i=1}^5 k_i \frac{a_i + \frac{a_i + c_i}{2} + c_i}{3} = \sum_{i=1}^5 k_i \frac{a_i + c_i}{2} = \sum_{i=1}^5 k_i b_i = b,$$

$$\text{or } X(M) = \frac{a + c}{2} = b \quad (3).$$

Therefore, for the defuzzification of M (a, b, c) one uses only its middle component $b = \sum_{i=1}^5 k_i b_i = b$.

Grey Numbers

Frequently in the everyday life, as well as in many applications of science and engineering, a system's data cannot be easily determined precisely and in practice estimates of them are used. The reason for this is that in large and complex systems, like the socio – economic, the biological ones, etc., many different and constantly changing factors are usually involved, the relationships among which are indeterminate, making their operation mechanisms to be not clear.

Nowadays two are the main tools for handling such approximate data: **Fuzzy Logic (FL)** (see Voskoglou, 2018e) which is based on the notion of *Fuzzy Set (FS)* initiated by Zadeh (1965), and the theory of **Grey System (GS)** initiated by Deng (1982). The GS theory has been mainly developed in China and has currently found important applications in agriculture, economy, management, industry, ecology, environment, meteorology, geography, geology, earthquakes, history, military affairs, sports, traffic, material science, biological protection and in many other fields of the human activity (see Deng, 1989 and its relevant references).

Roughly speaking, a GS is understood to be any investigated system with “poor” information. More explicitly, the systems which lack information, such as structure message, operation mechanism and behaviour document, are referred to as GSs. For example, the human body, agriculture, economy, etc., are GSs. On the grounds of existing grey relations and elements one usually can identify where “grey” means poor, incomplete, uncertain, etc.

The aim of the GS theory is to provide techniques, notions and ideas for analyzing latent and intricate systems, including the establishment of non-function models, the development of a grey process replacing an existing stochastic process, the transformation of disorderly raw data into a more regular series by grey generating techniques instead of modelling with the original data; grey decision making, grey forecasting control replacing classical control, the study of feeling and emotion functions and fields with whitening functions, the study in general of grey mathematics instead of classical mathematics, etc. (for more details see Deng, 1989).

An effective tool for handling the approximate data of a GS is the use of GNs. A GN is an indeterminate number whose probable range is known, but which has unknown position within its boundaries. The GNs are defined with the help of the real intervals. More explicitly, if $\mathbf{R}$ denotes the set of real numbers, a GN, say A , can be expressed mathematically by

$$A[a, b] = \{x \in \mathbf{R} : a \leq x \leq b\}.$$

If $a = b$, then A is called a **white number** and if $A(-\infty, +\infty)$, then it is called a **black number**.

Compared to the interval $[a, b]$, the GN $A[a, b]$ may enrich its uncertainty representation with a **whitening function** $g: [a, b] \rightarrow [0, 1]$ defining the **degree of greyness** $g(x)$, for each x in $[a, b]$. The closer is $g(x)$ to 1, the greater the probability of x to be the representative crisp value of the GN $A[a, b]$. For general facts on GNs we refer to Liu & Lin (2010).

The well known arithmetic of the real intervals (Moore, et al., 1995) has been used to define the basic arithmetic operations among the GNs. More explicitly, if $A[a_1, a_2]$ and $B[b_1, b_2]$ are given GNs and k is a real number, one defines:

- **Addition** by $A + B[a_1 + b_1, a_2 + b_2]$
- **Subtraction** by: $A - B = A + (-B)[a_1 - b_2, a_2 - b_1]$, where $-B[-b_2, -b_1]$ is defined to be the **opposite** of B .

- **Multiplication** by: $A \times B [min\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}, max\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}]$
- **Division** by: $A : B = A \times B^{-1} [min\{\frac{a_1}{b_1}, \frac{a_1}{b_2}, \frac{a_2}{b_1}, \frac{a_2}{b_2}\}, max\{\frac{a_1}{b_1}, \frac{a_1}{b_2}, \frac{a_2}{b_1}, \frac{a_2}{b_2}\}]$, where $0 \in [b_1, b_2]$ and $B^{-1} \in [\frac{1}{b_2}, \frac{1}{b_1}]$ is defined to be the **inverse** of B.
- **Scalar multiplication** by: $kA [ka_1, ka_2]$, if k is a positive real number and by $kA [ka_2, ka_1]$, if k is a negative real number.

Observe that $B + (-B) [b_1 - b_2, b_2 - b_1] \neq [0, 0] = 0$, $B + (-B) \neq (-B) + B \neq 0$ and $B \times B^{-1} = B^{-1} \times B \in [\frac{b_1}{b_2}, \frac{b_2}{b_1}] \neq [1, 1] = 1$.

The white number with the highest probability to be the representative real value of the GN $A[a, b]$ is denoted by $w(A)$. The technique of determining the value of $w(A)$ is called **whitenization** of A.

One usually defines $w(A) = (1-t)a + tb$, with t in $[0, 1]$. This is known as the **equal weight whitenization**. When the distribution of A is unknown (i.e. no whitenization function has been defined for A), one takes $t = \frac{1}{2}$, which gives that $w(A) = \frac{a+b}{2}$ (4).

Additional readings for all the topics covered in the present section are provided at the end of the Chapter.

FOCUS OF THE ARTICLE

Methodology

The remark which was stated in the Background Section, following the assessment of a system's mean performance using TFNs, gives the impulse to search for a "formal" assessment procedure, analogous to that with TFNs, that possibly reduces the required computational burden. This led us to utilizing GNs for the system's assessment.

For this, reconsidering the numerical scores A (100-85), B(84-75), C (74-60), D(59-50), F(49-0) attached to the corresponding qualitative grades in the above described method with the TFNs, one corresponds to each grade a GN, denoted for simplicity with the same letter, as follows: $A \in [85, 100]$, $B \in [75, 84]$, $C \in [60, 74]$, $D \in [50, 59]$ and $F \in [0, 49]$.

Then, by assigning to each of the system's components the GN corresponding to its individual performance and by using the same notation and analogous definitions as in the case of the TFNs, one finds that the **mean value** of all those GNs is equal to the GN:

$$M = \frac{1}{n} [n_A A + n_B B + n_C C + n_D D + n_F F] \quad (5)$$

The GN M can be considered as a grey representative of the system's mean performance. But $n_A A \in [85n_A, 100n_A]$, $n_B B \in [75n_B, 84n_B]$, $n_C C \in [60n_C, 74n_C]$, $n_D D \in [50n_D, 59n_D]$ and $n_F F \in [0n_F, 49n_F]$, therefore it turns out that

$M \in [m_1, m_2]$, with

$$m_1 = \frac{85n_A + 75n_B + 60n_C + 50n_D + 0n_F}{n} \text{ and } m_2 = \frac{100n_A + 84n_B + 74n_C + 59n_D + 49n_F}{n}.$$

Since the distributions of the GNs A, B, C, D and F are unknown, the same happens with the distribution of M. Therefore, one finds that

$$w(M) = \frac{m_1 + m_2}{2} \quad (6).$$

The value of $w(M)$ is the crisp outcome to be used for a quantitative estimation of the system's mean performance.

Remarks

1. Taking into account the way of construction of the TFNs and GNs A, B, C, D, F and on comparing the crisp outcomes obtained by the equations (3) and (6) defuzzifying / whitenizing their mean values $M(a, b, c)$ and

$M[m_1, m_2]$ respectively, it turns out that the assessment methods of a system's mean performance using as tools those TFNs and GNs respectively are **equivalent**, since they provide the same assessment outcomes.

2. Also, one observes that in the extreme case where the maximal possible numerical score corresponds to each component for each grade, i.e. the n_A scores corresponding to A are 100, the n_B scores corresponding to B are 84, etc., the mean value of all those scores is equal to b or m_2 respectively. Also, in the opposite extreme case, where the minimal possible numerical score corresponds to each student for each linguistic grade, i.e. the n_A scores corresponding to A are 85, the n_B scores corresponding to B are 75, etc. the mean value of all those scores is equal to a or m_1 respectively. Consequently, the assessment methods with the TFNs and the GNs give a **reliable approximation** of the system's mean performance and therefore they are useful when no numerical scores are used, but the system's performance is assessed by qualitative grades

SOLUTIONS AND RECOMMENDATIONS

In this section three applications to student, athlete and CBR computer systems' assessment are presented as examples for illustrating the use in practice and the equivalence of the two assessment methods with GNs and TFNs respectively and for comparing their outcomes to the classical assessment methods of calculating the mean value of scores and the GPA index. A fourth example follows that requires the solution of a system of equations with grey coefficients for assessing the performance of a football team in two different games

Example 1: The performance of four athletes was assessed by three different experts using a numerical scale from 0 – 100 as follows: A_1 (athlete 1): 48, 51, 52, A_2 : 87, 91, 93, A_3 : 82, 89, 94 and A_4 : 35, 40, 44. It is asked to evaluate the four athletes' mean performance.

Calculation of the mean value of the scores: The mean value of the $4 \times 3 = 12$ scores assigned to the athletes is approximately equal to 67.17, a score demonstrating a good (C) mean performance of the four athletes.

Solution using GNs: It can be observed that 5 of the scores assigned to the athletes correspond to the qualitative grade A, 1 corresponds to B, 2 to D and 4 scores correspond to F. Therefore, by assigning to the above qualitative grades the corresponding GNs one finds that their mean value is

$$M = \frac{1}{12} (5A+B+2D+4F) \in [50, 74.5].$$

Therefore $X(M) = \frac{50 + 74.5}{2} = 62.25$, which demonstrates again to a good (C) mean performance.

Solution using TFNs: In this case one finds the mean value

$$M = \frac{1}{12} (5A+B+2D+4F) = (50, 62.25, 74.5).$$

The defuzzification of M provides the same assessment outcomes with the method using the TFNs.

In concluding, when a system's performance is assessed with numerical scores, the calculation of the mean value of those scores gives the exact outcome for the assessment of its mean performance, whereas the methods using TFNs and GNs give an approximation of it. Therefore the last two methods are useful in practice only when the assessment is made by using qualitative grades.

Example 2: The following Table depicts the performance of two student groups, say G_1 and G_2 , in a common mathematical test involving the solution of three problems on the use of the derivatives for the maximization / minimization of a given function in one variable ².

It is asked to estimate and to compare the mean performance of the two groups in the test.

Table 1. Student performance

Grade	G_1	G_2
A	20	20
B	15	30
C	7	15
D	10	10
F	8	10
Total	60	85

Solution using GNs: Assigning to each student the corresponding GN and calculating the mean values M_1 and M_2 of those GNs for the groups G_1 and G_2 respectively one finds that:

$$M_1 = \frac{1}{60} (20A+15B+7C+10D+8F) \in [62.42, 79.33]$$

$$M_2 = \frac{1}{85} (20A+30B+15C+10D+10F) \in [62.94, 78.94]$$

Therefore the equation (6) gives that $w(M_1) \approx \frac{62.42 + 79.33}{2} \approx 70.88$ and $w(M_2) \approx \frac{62.94 + 78.94}{2} \approx 70.94$.

Consequently both groups demonstrated a good (C) mean performance, with the mean performance of the second group being slightly better.

Solution using TFNs: In this case the calculation of the mean values: gives that

$$M_1 = \frac{1}{60} (20A+15B+7C+10D+8F) = (62.42, 70.88, 79.33)$$

$$M_2 = \frac{1}{85} (20A+30B+15C+10D+10F) = (62.94, 70.94, 78.94)$$

Therefore, using the equation (3) for defuzzifying those mean values one obtains the same assessment outcomes as in the case of GNs. However, the use of TFNs requires the calculation of three components for obtaining the mean values in contrast to the use of GNs, which requires the calculation of two components only.

In concluding, although the above two methods are equivalent, the use of the GNs reduces the required computational burden.

Remark: The GPA index is calculated (Voskoglou, 2017, Chapter 6, p.125) by the formula

$$\text{GPA} = \frac{0n_F + 1n_D + 2n_C + 3n_B + 4n_A}{n} \quad (7).$$

Therefore, one finds that $\text{GPA} = \frac{4.20 + 3.15 + 2.7 + 1.10}{60} \approx 2.48$ for the first and $\text{GPA} = \frac{4.20 + 3.30 + 2.15 + 1.10}{85} \approx 2.47$ for the second group. Thus, in contrast to their mean performance, the first group demonstrated a slightly better quality performance than the second one. The above outcome illustrates the fact that the assessment of a system's quality performance could lead to a different outcome than the assessment of its mean performance.

Example 3 (Assessment of Case-Based Reasoning Systems):

Background: Case-Based Reasoning (CBR) is the process of solving new problems by adapting the solution of similar (analogous) problems solved in the past, which are usually referred as *past cases*. A case - library can be a powerful corporate resource allowing everyone in an organization to tap in it when handling a new problem. A CBR system, usually designed and functioning with the help of computers, allows the case-library to be developed incrementally, while its maintenance is relatively easy and can

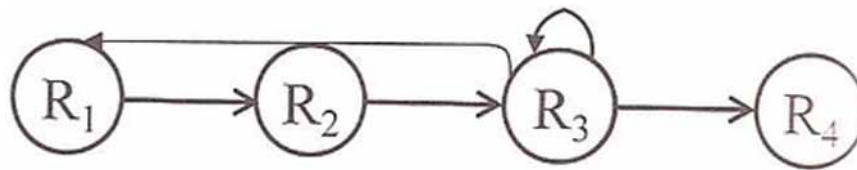
be carried out by domain experts. The CBR approach has got a lot of attention over the last 30-40 years, because as an intelligent – systems' method enables information managers to increase efficiency and reduce cost by substantially automating processes.

CBR has been formalized for purposes of computer and human reasoning as a four step process, often referred as the “four R’s”. These steps involve:

- R_1 : **Retrieve** the most similar to the new problem past case.
- R_2 : **Reuse** the information and knowledge of the retrieved case for the solution of the new problem.
- R_3 : **Revise** the proposed solution for solving the new problem.
- R_4 : **Retain** the part of this experience likely to be useful for future problem solving.

The first three of the above steps are not linear, characterized by a backward - forward flow among them. A simplified flow - chart of the CBR process, which is adequate for the purposes of the present paper, is presented in Figure 2:

Figure 2. A simplified flow-chart of the CBR process



More details about the CBR process and a detailed functional diagram illustrating the flow among its four steps are presented in (Voskoglou & Salem, 2014).

The Problem: Consider two CBR systems designed for help desk applications with their libraries containing 105 and 90 past cases respectively. Assume that the two systems' designers have supplied them with the same mechanism (software) that enables the assessment of the degree of success of each one

Table 2. Assessment of the past cases of the two CBR systems – First System

Steps	F	D	C	B	A
R_1	0	0	51	24	30
R_2	18	18	48	21	0
R_3	36	30	39	0	0

Table 3. Assessment of the past cases of the two CBR systems – Second System

Steps	F	D	C	B	A
R_1	0	18	45	27	0
R_2	18	24	48	0	0
R_3	36	27	27	0	0

of their past cases at each step of the CBR process, when used for the solution of new similar problems. Table 2 depicts the degree of success of their past cases in each of the three first steps of the CBR process

Here we shall compare the quality performance of the two systems by calculating the GPA index and their mean performance by applying our assessment methods with the TFNs and the GNs.

- i) **GPA index:** Denote by y_i , $i = 1, 2, 3, 4, 5$ the frequencies of the CBR system's cases whose performance is characterized by F, D, C, B and A respectively, i.e. $y_1 = \frac{n_F}{n}$, $y_2 = \frac{n_D}{n}$, etc. Then the equation (7) gives that

$$\text{GPA} = y_2 + 2y_3 + 3y_4 + 4y_5 \quad (8).$$

In case of the ideal performance ($y_5 = 1$, $y_4 = y_3 = y_2 = y_1 = 0$) we have $\text{GPA} = 4$, while in case of the worst performance ($y_1 = 1$, $y_5 = y_4 = y_3 = y_2 = 0$) we have $\text{GPA} = 0$; therefore $0 \leq \text{GPA} \leq 4$. Consequently, values of GPA greater than 2 could be considered as corresponding to a more than satisfactory system's performance. In our case, the data of Table 4 give the following frequencies:

Table 4. Frequencies of the past cases of the two CBR systems - First System

Steps	y_1	y_2	y_3	y_4	y_5
R_1	0	0	$\frac{51}{105}$	$\frac{24}{105}$	$\frac{30}{105}$
R_2	$\frac{18}{105}$	$\frac{18}{105}$	$\frac{48}{105}$	$\frac{21}{105}$	0
R_3	$\frac{36}{105}$	$\frac{30}{105}$	$\frac{39}{105}$	0	0

Table 5. Frequencies of the past cases of the two CBR systems – Second System

Steps	y_1	y_2	y_3	y_4	y_5
R_1	0	$\frac{18}{90}$	$\frac{45}{90}$	$\frac{27}{90}$	0
R_2	$\frac{18}{90}$	$\frac{24}{90}$	$\frac{48}{90}$	0	0
R_3	$\frac{36}{90}$	$\frac{27}{90}$	$\frac{27}{90}$	0	0

Replacing the values of frequencies from Table 3 in equation (8) one finds the following values for the GPA index:

$$\text{First System: } R_1: \frac{294}{105} = 2.8, R_2: \frac{177}{105} \approx 1.69, R_3: \frac{108}{105} \approx 1.03.$$

$$\text{Second System: } R_1: \frac{189}{90} = 2.1, R_2: \frac{168}{90} \approx 1.87, R_3: \frac{81}{90} = 0.9.$$

The above values of the GPA index show that the first system demonstrated a better quality performance at steps R_1 and R_3 (Retrieve, Revise), whereas the second one demonstrated a better performance at R_2 (Reuse). Further, the two systems' performance was proved to be more than satisfactory in R_1 and less than satisfactory in the other two steps, being worse at R_3 . This was logically expected, since the success in each step depends on the success in the previous steps. Notice that the two systems' performance at the last step R_4 was not examined, since ALL the past cases, even the unsuccessful ones, are retained in a system's library for possible use in future with related new problems; the unsuccessful ones to help for exploring possible reasons of failure to find a solution for the new problem.

Finally, the mean values of the GPA index for the two systems at the three steps R_1 , R_2 and R_3 are approximately equal to 1.84 and 1.62 respectively, showing that the first system demonstrated a better overall quality performance.

- ii) **Use of the TFNs:** From the data of Table 2 one finds that for the first system and in step R_1 we have 51 TFNs equal to $C(60, 67, 74)$, 24 TFNs equal to $B(75, 79.5, 85)$ and 30 TFNs equal to $A(85, 92.5, 100)$. The mean value of all those TFNs, denoted for simplicity by the same letter R_1 , is equal to

$$\begin{aligned} R_1 &= \frac{1}{105} (51C + 24B + 30A) = \frac{1}{105} [(3060, 3417, 3774) + (1800, 1908, 2016) + (2550, 2775, 3000)] \\ &= \frac{1}{105} (7410, 8100, 8790) \approx (70.57, 77.14, 83.71). \end{aligned}$$

Therefore, from equation (3) one gets that $X(R_1) = 77.14$, which shows that the first system demonstrated a very good (B) performance at step R_1 .

In the same way one calculates for the first system the mean values

$$R_2 = \frac{1}{105} (18F + 18D + 48C + 21B) \approx (51, 60.07, 69.14) \text{ and } R_3 = \frac{1}{105} (36F + 30D + 39C) \approx (36.57, 48.86, 61.14),$$

thus obtaining the analogous conclusions for the system's performance at the steps R_2 and R_3 of the CBR process.

Finally, the overall system's performance can be assessed by the mean value

$$R = \frac{1}{3} (R_1 + R_2 + R_3) \approx (52.71, 62.02, 71.33),$$

Therefore, since $X(R) = 62.02$, the system demonstrated a good (C) mean performance.

A similar argument gives for the second system the values $R_1 = (62.5, 68.25, 74)$, $R_2 \approx (45.33, 55.17, 65)$,

$R_3 = (33, .46.25, 59.5)$ and $R \approx (46.94, 56.56, 66.17)$, thus obtaining the analogous conclusions for its mean performance at each step of the CBR process and its overall mean performance.

- iii) **Use of the GNs:** According to this method for the first system and in step R_1 we have 51 GNs equal to $C \in [60, 74]$, 24 GNs equal to $B \in [75, 84]$ and 30 GNs equal to $A \in [85, 100]$. The mean value of all those GNs, denoted by R_1^* , is equal to

$$R_1^* = \frac{1}{105} (51C + 24B + 30A) \in [70.57, 83.71].$$

$$\text{Therefore, } W(R_1^*) = \frac{70.57 + 83.71}{2} = 77.14, \text{ etc.}$$

As we have seen in Section 4 this approach provides in general the same assessment outcomes with the use of TFNs, but, as it becomes more evident from the present example, it reduces significantly the required computational burden.

Table 6 depicts the assessment outcomes obtained in this example

Table 6. Outputs of the Assessment Methods Used in Example 3

METHOD	OUTPUT
GPA index	The first system demonstrated a better overall quality performance and a better quality performance at steps R_1 and R_3 , while the second system performed better at step R_2
TFNs / GNs	The first system demonstrated a better mean performance at all the steps of the CBR process

Example 4: The performance of a football team in two different games is represented by the GNs A and B respectively such that $A + B \in [2, 4]$ and $A \cdot B \in [1, 3]$. It is asked to compare the team's performance in the two games to its top possible performance corresponding to the numerical score 5

Solution: The whitenization of the given sum and product of the GNs A and B leads to the solution of the ordinary system $\begin{cases} x + y = 3 \\ x \cdot y = 2 \end{cases}$, which gives that $x = 2$, $y = 1$ or vice versa. Therefore $w(A) = 2$ and $w(B) = 1$ or vice versa, which means that the team's performance in the two games was equal to the 40% and to the 20% respectively of its top performance.

Remark: The attempt to solve the above problem by applying the definitions of addition and multiplication of GNs presented in the Background section leads to a system of equations with respect to the unknowns a_1, b_1, a_2, b_2 which does not provide a concrete solution. For more details on solving systems of equations with grey data see Voskoglou, 2018a.

FUTURE RESEARCH DIRECTIONS

Although the enormous development of technology during the last years makes easier and more comfortable the human life, it creates in parallel more and more complicated artificial systems, which are difficult to be managed by the traditional scientific methods. As a result, the applications of FL and of the GS theory have been rapidly expanded nowadays covering almost all the sectors of the human activities. In particular the FNs and the GNs have been proved to be efficient tools in handling approximate and uncertain data, playing an important role in fuzzy mathematics and in GS theory respectively, analogous to the role of the regular numbers in crisp mathematics.

Therefore, the attempt to use FNs or / and GNs as tools for solving other practical problems of the everyday life apart from the assessment ones (e.g. see Tanaka & Asai, 1984, Alevizos et al., 2017, etc.) looks as a very interesting and promising direction of future research on the subject. Those problems could be related to topics like Economics, Management, Human Cognition, Computer Science and Information Systems, Sports, Education, Natural Sciences, Psychology, Sociology and many others. It is of worth to notice that such efforts have been already started by the present author for solving equations, systems of equations and linear programming problems with fuzzy or grey data connected to real life applications (Voskoglou, 2018 a, b, c, d).

CONCLUSION

An inevitable consequence of the enormous technological development of the human society during the last years is the continuous development of more and more complicated artificial systems, whose function is characterized by vagueness and/ or uncertainty and it is frequently assessed by using approximate data. As a result the use of methods and principles of FL and of the GS theory have become recently very popular for treating, explaining and improving the performance of such kind of systems. The main tool for the application of the above two approaches is the use of the FNs and of the GNs respectively.

In the present work we have presented two innovative methods using TFNs and GNs for assessing a system's mean performance. Those methods are useful when qualitative grades and not numerical scores are used for the assessment purposes, because in this case the calculation of the mean value of those grades in the traditional way is impossible. Although our two methods were proved to be equivalent, the method with the GNs reduces significantly the required computational burden, since it requires the calculation of two components only (instead of three needed in the case of the TFNs) for obtaining the mean value in the form of a GN. Examples have been also presented on student, athlete, football teams and CBR systems assessment illustrating our methods in practice and showing that the system's quality performance, calculated by the traditional GPA index, may lead to different assessment outcomes. Future research directions on the subject have been also discussed.

REFERENCES

- Alevizos, P. D., Theodorou, Y., & Vrahatis, M. N. (2017). Correspondence Analysis with Grey Data: The Grey Eigen Value Problem. *Journal of Grey System*, 29(1), 92–105.
- Backlund, A. (2000). The definition of system. *Kybernetes*, 29(4), 444–451. doi:10.1108/03684920010322055

- Balley, K. D. (1994). *Sociology and the New Systems Theory: Toward a Theoretical Synthesis*. State of New York Press.
- Deng, J. (1982). Control Problems of Grey Systems. *Systems & Control Letters*, 288–294.
- Deng, J. (1989). Introduction to Grey System Theory. *Journal of Grey System*, 1, 1–24.
- Dubois, D., & Prade, H. (1980). *Fuzzy Sets and Systems: Theory and Applications*. Academic Press.
- Kaufmann, A., & Gupta, M. (1991). *Introduction to Fuzzy Arithmetic*. Van Nostrand Reinhold Company.
- Klir, G. J., & Folger, T. A. (1988). *Fuzzy Sets, Uncertainty and Information*. Prentice-Hall.
- Liu, S. F., & Lin, Y. (Eds.). (2010). *Advances in Grey System Research*. Springer. doi:10.1007/978-3-642-13938-3
- Moore, R. A., Kearfort, R. B., & Cloud, M. J. (1995). *Introduction to Interval Analysis*, 2nd Printing. SIAM.
- Tanaka, H., & Asai, K. (1984). Fuzzy Linear Programming Problems with Fuzz Numbers. *Fuzzy Sets and Systems*, 13(1), 1–10. doi:10.1016/0165-0114(84)90022-8
- van Broekhoven, E., & De Baets, B. (2006). Fast and accurate centre of gravity defuzzification of fuzzy systems outputs defined on trapezoidal fuzzy partitions. *Fuzzy Sets and Systems*, 157(7), 904–918. doi:10.1016/j.fss.2005.11.005
- Voskoglou, M. Gr. & Salem A.-B.M. (2014). Analogy-Based and Case-Based Reasoning: Two Sides of the Same Coin. *International Journal of Applications of Fuzzy Sets and Artificial Intelligence*, 4, 7–18.
- Voskoglou, M. G. & Theodorou, Y. (2017). Application of Grey Numbers to Assessment Processes. *International Journal of Applications of Fuzzy Sets and Artificial Intelligence*, 7, 273–280.
- Voskoglou, M. G. (2017a). Application of Fuzzy Numbers for Assessment Processes. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (4th ed., pp. 3216–3225).
- Voskoglou, M. G. (2017b). *Finite Markov Chain and Fuzzy Logic Assessment Models: Emerging Research and Opportunities*. Amazon.
- Voskoglou, M. G. (2018a). Solving Systems of Equations with Grey Data. *International Journal of Applications of Fuzzy Sets and Artificial Intelligence*, 8, 103–111.
- Voskoglou, M. G. (2018b). Solving Linear Programming Problems with Grey Data. *Oriental Journal of Physical Sciences*, 3(1), 17–23. doi:10.13005/OJPS03.01.04
- Voskoglou, M. G. (2018c). (to appear). Fuzzy Linear Programming. *Egyptian Journal of Computer Science*, 42(3).
- Voskoglou, M. G. (2018d). Solving Systems of Equations with Fuzzy Data. *The Journal of Physiological Sciences; JPS*.
- Voskoglou, M. G. (2018e). Fuzzy Logic: History, Methodology and Applications to Education, *Sumerian Journal of Education. Linguistics and Literature*, 1(1), 10–18.
- Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, 8(3), 338–353. doi:10.1016/S0019-9958(65)90241-X

ADDITIONAL READING

Huang, G., & Dan Moore, R. (1993). Grey linear programming, its solving approach, and its application. *International Journal of Systems Science*, 24(1), 159–172. doi:10.1080/00207729308949477

Verdegay, J. L. (1984). A dual approach to solve the fuzzy linear programming problem. *Fuzzy Sets and Systems*, 14(2), 131–141. doi:10.1016/0165-0114(84)90096-4

Voskoglou, M. G. (2011). *Stochastic and Fuzzy Models in Management, Education and Artificial Intelligence*, Saarbrücken. Lambert Academic Publishing.

Voskoglou, M. G. (2015). Defuzzification of fuzzy numbers for student assessment. *American Journal of Applied Mathematics and Statistics*, 3(5), 206–210.

KEY TERMS AND DEFINITIONS

Defuzzification: The fuzzy logic's approach for a problem's solution involves the fuzzification of the problem's data by representing them with properly defined fuzzy sets, the evaluation of the fuzzy data in order to express the problem's solution in the form of a unique fuzzy set and the defuzzification of this fuzzy set in order to translate the problem's mathematical solution to the natural language for use with the original real-life problem. The most popular defuzzification method is the Centre of Gravity (COG) Technique in which the corresponding system's fuzzy outputs are represented by the coordinates of the COG of the level's section contained between the graph of the membership function expressing those outputs and the OX axis.

Fuzzy Logic (FL): A logic based on the concept of fuzzy set that, in contrast to the classical bi-valued logic of Aristotle (yes – no), characterizes cases with multiple values.

Fuzzy Number (FN): A fuzzy set A on the set R of the real numbers which is normal (i.e. there exists x in R such that $m_A(x) = 1$) and convex (i.e. all its a -cuts $A^a = \{x \in U: m_A(x) \geq a\}$, a in $[0, 1]$, are closed real intervals) and whose membership function $y = m_A(x)$ is a piecewise continuous function.

Fuzzy Set (FS): A generalization of the concept of crisp set that gives a mathematical formulation to real situations in which certain definitions have not clear boundaries (e.g. the high mountains of a country).. A FS, say A , is characterized by its membership function $y = m_A(x)$ defined on the universal set of the discourse U and taking values in the interval $[0, 1]$, thus assigning to each element x of U a membership degree $m_A(x)$ with respect to A . The closer is $m_A(x)$ to 1, the better x satisfies the characteristic property of A .

Grade Point Average (GPA): A weighted average assigning greater coefficients (weights) to the higher scores and therefore focusing on a system's quality performance.

Grey Number (GN): An indeterminate number whose probable range is known, but which has unknown position within its boundaries. If R denotes the set of real numbers, a GN, say A , can be expressed mathematically with the help of a real interval $[a, b]$ in the form $A[a, b] = \{x \in R : a \leq x \leq b\}$. If $a = b$, then A is called a white number and if $A \in (-\infty, +\infty)$, then it is called a black number. Compared with the interval.

Grey System (GS): A system which lacks information, such as structure message, operation mechanism and behaviour document. Usually, on the grounds of existing grey relations and elements one can identify where “grey” means poor, incomplete, uncertain, etc.

System (in Management): A set of interacting components forming an integrated whole and working together for achieving a common target. A factory, a hospital, or a bank constitute common examples of systems.

Triangular Fuzzy Number (TFN): The simplest form of a FN, whose membership function's graph forms a triangle with the OX-axis. The defuzzification of a TFN $A = (a, b, c)$ may be obtained by considering as its representative crisp value the x-coordinate $X(A) = \frac{a+b+c}{3}$ of the COG of its membership function.

Whitenization of a gn: The process of determining the white number with the highest probability to be the representative real value of a GN $A[a, b]$, which is denoted by $w(A)$. When the distribution of A is unknown, one usually takes $w(A) = \frac{a+b}{2}$.

ENDNOTES

- ¹ One could add more qualitative grades for the assessment, e.g. E = less than satisfactory, could be added between D and F, etc. Also the assignment of the numerical scores to each grade is not uniquely made, depending on the user's personal goals. For example, in a more strict assessment one could take A (90-100), B(80-89), C (70-79), D (60-69), F (0-59), etc. However, the above alternatives do not change the generality of the assessment method proposed.

- ² The following problems were given for solution to the two student groups in the classroom application:

Problem 1: One wants to construct a channel to run water by bending across to its longest side the two edges of a rectangle metallic leaf with dimensions 20 cm and 32 cm respectively, so that to be perpendicular to the other part of the leaf. Assuming that the flow of water is constant, how we can run the maximum possible quantity of the water through the channel?

Problem 2: Among all the cylindrical towers having a total surface of $180\pi \text{ m}^2$, which one has the maximal volume?

Problem 3: A car dealer has a mean annual demand of 250 cars. The annual cost of storing a car is 100 euro and each time he makes a new order he pays an extra amount of 2000 euro for general expenses (transportation, insurance etc). It has been arranged for the cars of the new order to arrive at the time when the last car of the previous order has been sold. How many cars must be ordered each time to minimize the total cost for the dealer?

Section 3

Cyber Crime, Cyber Bullying, and Digital Terrorism

Cyber Profiling in Criminal Investigation

3

Szde Yu

Wichita State University, USA

INTRODUCTION

As technologies are progressively ingrained in modern life, the line between street crimes and cybercrimes has become gradually blurry. Nowadays, in almost all criminal investigations there is a need for digital investigation to be incorporated as digital information has become relevant and even crucial in most cases. Investigators often have to rely on information retrieved from electronic platforms to find leads and evidence even in crimes that would be traditionally considered street crimes. When comprehensive information is not always attainable, the use of cyber profiling may provide further insight into crucial questions regarding identity, time, and location. Cyber profiling is an analysis on the digital footprints associated with a person whose identity may or may not have been known. When the identity is unknown, the purpose of cyber profiling is primarily about identifying this person or at least creating a workable profile that narrows down suspects. When the identity is already known, the purpose is usually aimed to reveal information that is not readily visible, such as personality, motives, or linkage to other crimes. Besides being subject-centered, cyber profiling can also be event-centered in an attempt to identify time and location. As such, cyber profiling has a wide variety of application in criminal investigation and yet it has not been duly valued. This chapter introduces the fundamentals of cyber profiling and calls for more attention to the development of this technique.

BACKGROUND

To date, there is very little research effort in developing or validating cyber profiling. Therefore, it could be easy to dismiss cyber profiling. Yet, cyber profiling actually has been widely applied in many fields, albeit the name might not be recognized. In other words, many people are already experiencing cyber profiling without knowing it. For instance, some experts tend to make diagnosis about a public figure's personality or state of mind based on comments made on Twitter or other social media. This is basically a form of cyber profiling as the inference is based on digital footprints in the absence of any physical contact. When people do online shopping on certain products and consequently receive promotional advertisements about similar products, it is another example of cyber profiling because the online vendors are profiling shoppers based on their past shopping habits so as to decide what other products they would be interested in as well. Online dating services like Tinder also allows users to do cyber profiling based on only a few photos before they decide which stranger they want to go out on a date with. If cyber profiling can be relied on in the business world and social media, there is no reason why it cannot be utilized in criminal investigations.

DOI: 10.4018/978-1-7998-3479-3.ch024

Currently, criminal investigators are already used to looking for clues in a subject's digital footprints, such as text messages, emails, social media postings, and online shopping or browsing records (Rogers, 2003). However, cyber profiling calls for more attention to the implicit information hidden in the digital footprints. For instance, a man might intentionally avoid expressing his political view on the Internet, but from an analysis on the videos he watched on YouTube and the news websites he frequently visited we can still deduce what this man's political stance is, provided access to such digital footprints is not an issue. The same analysis might shed light on other aspects as well. A study was conducted on Facebook to test the reliability of cyber profiling (Yu, 2013), and it found that using nothing but the participant's Facebook public information, it is possible for a trained profiler to draw largely accurate conclusions about the participant's race, age, gender, and nationality. Moreover, profilers may correctly infer a person's certain personality traits, such as self-control based solely on digital footprints. Lambiotte and Kosinski (2014) also reported the ability to predict personality based on publicly available digital footprints, and they believed such prediction will become more and more accurate as more electronic data are being generated on a daily basis. This type of prediction, aka cyber profiling, can apply to criminal investigation and become a huge help.

Being able to predict personality and characteristics like gender and age has always been an important aspect of criminal profiling (Williams, Nathanson, & Paulhus, 2010; Douglas & Burgess, 1986; Pinizzotto & Finkel, 1990; Canter, 2004). While traditional criminal profiling focuses on behavioral evidence (Turvey, 2011), it becomes crippled when there is no physical trace serving as behavioral evidence. It is a serious issue in modern day investigations where many social interactions only take place in the virtual world (Rogers, 2003). In this regard, cyber profiling uses digital footprints as online behavioral evidence to look for information that can identify characteristics and personality (Yu, 2013). If identifying information can be directly retrieved, then profiling is not needed. For example, it should be easy to tell a person's gender and race by simply looking at the person's selfies. Profiling, however, becomes crucial when such identifying information is vague or deliberately disguised. For instance, when a 46-year-old man pretends to be a 16-year-old girl on the Internet, the photos shared by him probably should not be trusted. In an era where almost everyone has some sort of online presence, it is imperative for criminal investigators to possess the skills and knowledge to verify the authenticity of each online persona they encounter in the course of investigation. This is where cyber profiling begins. Further, from the online persona's digital footprints, investigators should develop skills to find useful clues and evidence that can point to the right person and the right place. Moreover, investigators ought to be able to recognize potential linkage among the digital footprints generated by different online personas in case they are actually owned by the same person. For example, a person can own multiple accounts on Facebook or Twitter and use each account to play a different role in cyberspace. Can the investigators see through this and make connections? This is where cyber profiling helps.

Unfortunately, cyber profiling is not yet a technique that law enforcement agencies are concerned with, even though their agents constantly have to look for clues in a suspect's or a victim's digital footprints. Most of the time, investigators only see what is there to be seen, and tend to overlook what is hidden or implicit in the digital footprints. For example, some researchers found that criminals could easily disguise incriminating communication as email spam and investigators would likely overlook it even when the disguise is very primitive (Yu, 2015). In this regard, cyber profiling may help reveal what is not so obvious to be seen considering incriminating information is usually implicit.

FOCUS OF THE ARTICLE

The Goal of Cyber Profiling

The concept of cyber profiling is derived from criminal profiling. The primary goal of criminal profiling is to provide the criminal investigator with a psychological and social assessment of the subject (Holmes & Holmes, 2009). Cyber profiling shares the same goal, but it relies on the subject's digital footprints instead of physical behavior to perform a psychological and social assessment that helps reveal the subject's identity or criminality. Cyber profiling can be regarded as an extension of criminal profiling, so most principles of criminal profiling still apply in terms of cyber profiling. The following sections discuss the three fundamental elements that constitute cyber profiling, including online behavior analysis, victimology, and cyber-context analysis (Yu, 2013). These three elements are derived from the three elements of behavioral evidence analysis, including forensic analysis, victimology, and crime scene analysis (Turvey, 2011). Behavioral evidence analysis is a profiling approach that focuses on physical behavior, while in the following this focus is adjusted to fit the digital environment where cyber profiling is called for.

Online Behavior Analysis

Any behavior that generates observable digital footprints is considered online behavior (Yu, 2014). For instance, when someone posts a photo on social media, the act of posting the photo is online behavior and the photo is the digital footprint. Even when this photo is deleted, the removal of this photo should be seen as online behavior as well. Accordingly, online behavior analysis focuses on linking digital footprints that are associated with the same person and examine the content of these digital footprints to find identifying information as previously discussed. In addition to the content, online behavior analysis must also examine the rationale behind the online behavior.

For instance, a person's choosing of a password is a form of online behavior, but the reason why this particular password was chosen also matters. If the rationale behind this password choice can be figured out, the password might be correctly disclosed and the rationale can also help reveal some of the subject's thinking patterns and logic. Besides, just because people manifest the same online behavior, it does not necessarily mean their behavior can be explained by the same rationale. For example, some people use Facebook for social-networking, while some might be using it for research only (Yu, 2014).

Moreover, the dynamics of the online behavior is another focal concern in online behavior analysis. Human behavior is not a constant. It may evolve or devolve over time. It may also manifest in response to the environmental restrictions rather than to free will (Yu, 2013; Yu, 2014). Therefore, it is wrong to assume it is not the same person simply because the online behavior changes. Likewise, it is inappropriate to assume it must be the same person simply because the online behavior appears to be the same. Especially on electronic platforms, many things have to be done in a certain way rather than according to personal preferences. For example, some social media sites do not allow users to change the background color, and therefore it is not suitable to use the background color to profile the user's preference. However, to the extent where users can indeed make customized changes, such as profile photos or privacy settings, the changes over time may be worth analyzing because they could be indicative of some personal evolvement, such as mood changes, perception changes, or increased knowledge (Yu, 2014).

To sum up, online behavior analysis should be concerned with three aspects: the content of the behavior, the rationale behind the behavior, and the dynamics of the behavior. Conceivably, the digital footprints attainable may not always be sufficient to support analysis on all three aspects. In such cases, investigators or profilers need to be cautious about any conclusions drawn from cyber profiling. In fact, cyber profiling is aimed to provide directions rather than conclusions. It is an aid, not a solution.

Victimology

In addition to online behavior analysis, similar to criminal profiling, cyber profiling also involves victimology (Turvey, 2011). Victimology is a study of the victim's role in a criminal event as the victim's characteristics and actions/reactions might have directly or indirectly account for the outcome of victimization. Cyber profiling looks into victimology to understand how certain people became victims in an online setting while others in the same settings did not (Yu, 2013). For example, why do some computers become hijacked by ransomware, while others in the same organization do not? Why do some people become a victim of cyber-bullying while others are not being targeted? Questions of this sort are the center of victimology. To be clear, this is not an attempt to blame the victim for victimization. From an investigative perspective, this is a crucial step toward understanding how the victim was chosen, and in turn such understanding may shed more light on how the perpetrator operates and even where they operate (Turvey, 2011).

In criminal investigations, there are three important elements to be determined: motive, means, and opportunity. The motive of a criminal act is not always clear and at times it can be misjudged due to misleading clues. For instance, when a man is killed and his wallet is missing, the motive could be easily assumed to be about money. However, it could be a hate crime at first and the robbery is simply incidental. This is why profiling is needed, because profiling is intended to see beyond what things appear to be. Cyber profiling has the same intention and studying victimology can help reveal all possible motives, not just the ostensible one. Studying the victim also helps reveals the perpetrator's means or capability. Not everyone is capable of committing a crime, especially when it requires advanced computer skills and equipment. Studying victimology can help determine what level of skills and resources this perpetrator may possess. Finally, opportunity is a necessary condition in all crimes. Some criminals have strong motives and great skills, but without opportunity they still cannot make things happen. For instance, a hacker cannot steal money from your bank account if you do not own a bank account. Therefore, cyber profiling studies the victim to see how the victim might have inadvertently provided opportunities for the perpetrator, and more importantly how the perpetrator exploited these opportunities. The way criminals take advantage of opportunities or even the ability to create new opportunities can usually further delineate their means and motives.

In conclusion, cyber profiling studies victims to help verify motive, means, and opportunity. In other words, cyber profiling analyzes more than just the criminal's digital footprints as it analyzes the victim's digital footprints as well. By studying multiple victims, it can detect potential linkage among different incidents even when the criminal does not leave behind too many digital footprints to be profiled.

Cyber-Context Analysis

After analyzing online behavior and victimology, cyber profiling has to also take into account the context in which online behavior takes place. More often than not, human behavior is a reaction to the context, and online behavior is no exception. The same behavior in different contexts might bear very different mean-

ings. Anything taken out of context could be massively misinterpreted. When it comes to cyber-context, it is important to be mindful that online behavior is all computer-mediated. This follows that online behavior is unlikely to be a spontaneous or instinctive response. For example, when a man is enraged, he might slam a door or throw a chair, and that behavior can be seen as a spontaneous response driven by emotion. The behavior probably is not a result of rational thinking. In contrast, for him to go on the Internet to express his anger by sending some angry messages, he must exercise his rational mind to some extent before he can accomplish such online behavior. The fact that all online behavior is computer-mediated makes cyber-context more artificial and contrived. Hence, in cyber profiling it is important to understand the online settings where online behavior and online interaction occur because these settings might be either facilitating or suppressing certain expression of a person's true intent. For instance, typing "LOL" in an online chatroom does not necessarily indicate real laughter. It could be simply seen as an online expression that seems appropriate in a particular cyber-context even when the person does not feel like laughing at all.

In addition to being computer-mediated, many cyber-contexts tend to provide a sense of anonymity. It does not have to be really anonymous. As long as online users feel their true identity is obscure enough, this sense of anonymity tends to result in lowered inhibition that could potentially lead to more deviance (Joinson, 1999; Kabay, 1998; Suler & Phillips, 1998). The connotation is that in such seemingly anonymous contexts, people might be more prone to express their true opinions or to reveal the dark side of their personality (Yu, 2013). As a result, profilers might notice inconsistent patterns between a subject's online behavior and daily behavior. On the other hand, when anonymity is not so assured, most people tend to have a heightened self-awareness (Gonzalez & Hancock, 2011), which could lead to an attempt to hide their true feelings or to manipulate the observer's perception as a way of self-promotion (Gonzalez & Hancock, 2011; Mehdizadeh, 2010; Stefanone et al., 2011; Wilson et al., 2010). Consequently, online behavior in these cyber-contexts is more likely to reveal what people want the audience to see instead of what they really are. Either way, profilers cannot just look at online behavior and assume it is always a truthful reflection of the subject's personality or genuine thought. Hence, the cyber-context must be analyzed as part of cyber profiling to take into account this audience effect.

The Issue

Similar to criminal profiling, cyber profiling requires ample behavioral traces to be reliable. Without abundant digital footprints associated with the subject, profiling might turn into a guessing game based on nothing but the profiler's presumptuous opinions. Therefore, access to relevant digital footprints is vital. Fortunately, nowadays many people, especially the younger generation, seem to be fond of sharing some aspects of their private life on the Internet to the public. When there is a probable cause, it is also not too difficult to obtain a court order for investigators to gain access to less public information, such as online communication and file storage. However, it may not be enough as most people's digital life is mostly private. Simply put, you probably do not want people to know about most of the things you do on your phone or computer, for whatever reasons. Hence, the ability to uncover all hidden digital footprints is a big challenge. Another frequently encountered problem is to correctly associate digital footprints with the right person. Sometimes the investigators have no problem gaining digital footprints from a variety of sources but do not know with whom these footprints should be associated. As a result, the investigators could be using only a fraction of the usable materials due to oversight or they could be mixing digital footprints from different contributors. In this regard, someone with expertise in cyber profiling can be of tremendous assistance. Regrettably, such expertise is not being emphasized as most police agencies tend to see only computer forensics as what they need in a digital investigation.

Cyber profiling is not in competition with computer forensics. In fact, it complements computer forensics. Some even argue cyber profiling should be seen as an integral component of computer forensics (Shaw, 2006; Yu, 2013). Some aspects of cyber profiling actually need to be determined by means of computer forensics, such as IP addresses and systemic logs (Steel, 2014). Moreover, cyber profiling can help identify more digital evidence for forensic analysis, and it can help interpret digital evidence and make inference. While computer forensics is about using generalized rules to analyze digital footprints, cyber profiling is looking into the unique circumstances in each case by examining online behavior, victimology, and cyber-context. Again, they are complementary to each other rather than in competition.

Needless to say, it requires a special skill set to analyze digital footprints adequately. The following section outlines the proficiency needed for a qualified cyber profiler.

SOLUTIONS AND RECOMMENDATIONS

Cyber profilers need to be proficient in a few areas, including computer forensics, criminology, Internet settings and cultures, text analytics, cyber-psychology, and reasoning skills. It could be hard for one person to be proficient in all areas as they require different types of training, so cyber profiling should be performed as teamwork if possible. However, a profiler needs to at least dabble at all areas and build some background knowledge in each one of them. For example, if an expert in computer forensics lacks reasoning skills, he or she probably will draw wrong conclusions from the forensic reports. Likewise, a linguist who is unfamiliar with the Internet cultures will be prone to misinterpret the textual interaction in an online setting. More importantly, cyber profilers should always be mindful of the different languages and ethnic cultures that could be easily integrated or intertwined on the Internet, which is much more discernible in the physical world. For example, a Chinese man might rarely speak English in a face-to-face conversation, but could be used to typing English on the Internet simply because it is faster than typing Chinese characters.

While cyber profiling requires expertise from many disciplines built on different theories, some basic principles apply to all of them. These principles are once again derived and adapted from some principles related to criminal profiling.

1. The principle of uniqueness. Every case is unique as no two individuals will develop in exact same fashion and result in the same behavior (Turvey, 2011). The individual context needs to be analyzed separately in each case before behavior can be interpreted accurately. Experience is invaluable but trying to find every answer from the past experience may not work well with a new problem. Especially in terms of cyber profiling, as technologies rapidly advance, old perspectives may not apply to the new case anymore.
2. The principle of behavioral dynamics. People's behavior could evolve or devolve so they do not remain unchanged over time, even when the motive is the same (Turvey, 2011). People change. Sometimes these changes could actually be more noticeable in online behavior, if the behavioral evidence has been preserved over time. Not only can the online behavior change over time, the cyber-context could be changing as well and thus account for behavioral changes.
3. The principle of motivation. All behaviors have underlying motivation, although sometimes the motivation could be subconscious (Turvey, 2011). Profilers should endeavor to uncover the motivation. Online behavior can be complex and one behavior can serve more than one purpose. For instance, an email spammer could use a fake server to cover traces and to bypass spam filters at the

same time (Yu, 2011; Yu, 2015). Conversely, multiple behaviors could be motivated by the same goal. For example, in cyberspace people could be doing different things with only one purpose in mind: to attract more attention. Furthermore, different people could engage in the same behavior for different reasons. One simple example is that while many people can click on the “like” button on Facebook, it does not mean they all “like” it for the same reason (Yu, 2014). This is why analyzing the rationale behind behavior is crucial in online behavior analysis, as previously mentioned.

4. The principle of motivational dynamics. The motive could change during the act (Turvey, 2011). This is especially true when the behavior involves interaction. For example, what started as friendly bantering online could turn into cyberbullying. Usually when the motivation has changed, the ensuing online behavior will change accordingly. However, such changes may not always be evident, especially when the subject deliberately makes an effort to conceal the motivational change. In this case, to better detect motivational dynamics, online behavior needs to be examined in the context as a sequence of events rather than isolated events.
5. The principle of unintended consequences. Behavior could be an unintended result (Turvey, 2011). In a cyber-context, a behavior could easily result from technical unfamiliarity or mistakes, rather than the true intent. This is why sometimes people send a blank email by accident, or reveal sensitive information online without knowing it. Moreover, even with the best intent, people could still provide misinformation, such as a clerical error. This is why it is important for profilers to analyze online behavior in the context.
6. The principle of humanity. While it is easy to focus on the technological aspects in a digital investigation, crime is still essentially a human problem (Yu, 2013). Criminals are not expected to be forthcoming, and this dishonest tendency can affect their digital footprints as well. Even good people still have reasons to be untruthful from time to time, but at the same time some people are actually more likely to show their true colors in an online environment (Suler & Phillips, 1998; Yu, 2015). To better access the human element in each case, cyber profilers need to have a good understanding on human mind and criminality in addition to computer knowledge. Since there is no border on the Internet, the cultural factors are important in the understanding of human mind (Yu, 2013).
7. The principle of subjectivity. Profiling in any form is inevitably subjective (Yu, 2013). Being subjective is not necessarily a bad thing. Many experts are better at their job precisely because their subjective view is superior to that of the ordinary people. Nevertheless, a good profiler should always entertain the possibility of being wrong and avoid succumbing to common logical fallacies or arrogance. It is also imperative for cyber profilers to remember that their subjective perception may not coincide with the perception of their suspect (Yu, 2015). In other words, what makes sense to them may not make sense to you and vice versa. The suspect’s perception of reality should take precedence over the profiler’s subjective perception because the suspect, not the profiler, is the one responsible for the digital footprints being analyzed.
8. The principle of linkage. When dealing with digital footprints, cyber profilers need to be able to discern the source (Yu, 2013; Yu, 2015). If a computer was shared by more than one person, the digital footprints extracted from this computer may have multiple contributors. Conversely, a person may use more than one device or account, and hence leave behind digital footprints on multiple sources. The ability to correctly link digital footprints to the rightful contributor is of the foremost importance to a cyber profiler. False positives would attribute irrelevant digital footprints to a wrong contributor and render cyber profiling unreliable. False negatives could result in an oversight on critical digital footprints and thus make cyber profiling unreliable as well.

FUTURE RESEARCH DIRECTIONS

Cyber profiling or digital footprints analysis deserves more attention in the field of criminal investigation. It is a technique not only helpful but also necessary, given how often crime involves digital footprints nowadays. Future research has two primary goals. First, how to effectively apply cyber profiling in various types of crime needs to be studied. This involves assessing the nature of the crime, identifying relevant digital footprints, acquiring access to such digital footprints, connecting digital footprints to the correct contributor, and producing reliable evidence. Second, research must look into how to perform adequate analysis in a systematic manner so that it is easier to train profilers. When profiling becomes nothing but the profiler's hunch, it is difficult to teach anyone else how to do it. On the other hand, profiling cannot be done by simply following a list of steps, because profiling must maintain flexibility to adjust to the unique circumstances in every case. Put differently, it is unrealistic to expect every profiler can produce the same results by following the same instructions, and it is equally unrealistic to expect the same instructions will suffice in every case. In this regard, research ought to build on what has been studied about criminal profiling and extend it to cyber profiling by recognizing their similarity as well as discrepancy. In this chapter, the discussion of cyber profiling is largely based on the framework of behavioral evidence analysis (Turvey, 2011), but future research might look into other frameworks that take into account more behavioral features in an online setting.

CONCLUSION

Cyber profiling performs an idiographic analysis on a subject's digital footprints which result from the subject's online behavior. It is not generally aimed to create generalizable rules or statistics, although this aim is not entirely objectionable. The primary focus is generally on the particular subject in question. The subject can be a suspect, a victim, or sometimes just a person of interest. The identity of this subject could be known or unknown. Either way, cyber profiling may offer insight into this subject's background, characteristics, and personality. From this, other information may be further inferred, such as whereabouts, personal history, or anything else that helps solve a crime. Cyber profiling can also be event-centered and used to link separate incidents together when the same person is responsible for them. Since cyber profiling relies on digital footprints, the quality and quantity of digital footprints are equally important. Irrelevant (i.e., false positive) or insufficient (i.e., false negative) digital footprints can both be detrimental to profiling. Therefore, the first step of cyber profiling always involves discerning and accessing applicable sources of online behavior, followed by a thorough online behavior analysis, victimology study, and cyber-context analysis, which may in turn point to more useable digital footprints. In addition to the content of online behavior, the rationale and the dynamics of online behavior are also the focal concerns in cyber profiling. Although they may not always be directly visible, this is actually why profiling is needed. Profiling is about reading between the lines, rather than merely describing what can be seen by everyone. Only with a comprehensive understanding of online behavior, victimology, and cyber-context can the resulting digital footprints be correctly interpreted. Cyber profilers may have expertise in diverse areas, but the basic principles apply to all.

REFERENCES

- Canter, D. (2004). Offender profiling and investigative psychology. *Journal of Investigative Psychology and Offender Profiling*, 1(1), 1–15. doi:10.1002/jip.7
- Douglas, J. E., & Burgess, A. E. (1986). Criminal profiling: A viable investigative tool against violent crime. *FBI Law Enforcement Bulletin*, 55, 9–13.
- Gonzalez, A., & Hancock, J. (2011). Mirror, mirror on my Facebook wall: Effects of exposure to Facebook on self-esteem. *Cyberpsychology, Behavior, and Social Networking*, 14(1-2), 79–83. doi:10.1089/cyber.2009.0411 PMID:21329447
- Holmes, R. M., & Holmes, S. T. (2009). *Profiling violent crimes: An investigative tool*. SAGE Publication Inc.
- Joinson, A. (1999). *Social desirability, anonymity, and internet-based questionnaires*. Academic Press.
- Kabay, M. (1998). *Anonymity and pseudonymity in cyberspace: Deindividuation, incivility and lawlessness versus freedom and privacy*. Available at: http://www.cs.wright.edu/~pmateti/InternetSecurity/Lectures/Privacy/Anonymity_Pseudonymity.PDF
- Lambiotte, R., & Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12), 1934–1939. doi:10.1109/JPROC.2014.2359054
- Pinizzotto, A. J., & Finkel, N. J. (1990). Criminal personality profiling: An outcome and process study. *Law and Human Behavior*, 14(3), 215–233. doi:10.1007/BF01352750
- Rogers, M. (2003). The role of criminal profiling in the computer forensics process. *Computers & Security*, 22(4), 292–298. doi:10.1016/S0167-4048(03)00405-X
- Shaw, E. D. (2006). The role of behavioral research and profiling in malicious cyber insider investigations. *Digital Investigation*, 3(1), 20–31. doi:10.1016/j.diin.2006.01.006
- Steel, C. M. (2014). Idiographic Digital Profiling: Behavioral Analysis Based On Digital Forensics. *The Journal of Digital Forensics. Security and Law: JDFSL*, 9(1), 7–18.
- Stefanone, M. A., Lackaff, D., & Rosen, D. (2011, January). Contingencies of self-worth and social networking site behavior. *Cyberpsychology, Behavior, and Social Networking*, 14(1-2), 41–49. doi:10.1089/cyber.2010.0049
- Suler, J. R., & Phillips, W. L. (1998). The bad boys of cyberspace. *Deviant Behavior*.
- Turvey, B. E. (2011). *Criminal profiling: An introduction to behavioral evidence analysis*. Academic Press.
- Williams, K. M., Nathanson, C., & Paulhus, D. L. (2010). Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation. *Journal of Experimental Psychology. Applied*, 16(3), 293–307. doi:10.1037/a0020773 PMID:20853988
- Wilson, K., Fornasier, S., & White, K. M. (2010, April). Psychological predictors of young adults' use of social networking sites. *Cyberpsychology, Behavior, and Social Networking*, 13(2), 173–177. doi:10.1089/cyber.2009.0094

Yu, S. (2011). Email spam and the CAN-SPAM Act. *International Journal of Cyber Criminology*, 5(1), 715–735.

Yu, S. (2013). Behavioral evidence analysis on Facebook: A test of cyber-profiling. *Defendology*, 16(33), 17–30. doi:10.5570/dfnd.en.1333.02

Yu, S. (2014). Does low self-control explain voluntary disclosure of personal information on the Internet? *Computers in Human Behavior*, 37, 210–215. doi:10.1016/j.chb.2014.04.055

Yu, S. (2015). Covert communication by means of email spam: A challenge to digital investigation. *Digital Investigation*, 13, 72–79. doi:10.1016/j.diin.2015.04.003

ADDITIONAL READING

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40(1), 129–152. doi:10.1146/annurev-soc-071913-043145

Lambiotte, R., & Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12), 1934–1939. doi:10.1109/JPROC.2014.2359054

Shaw, E. D. (2006). The role of behavioral research and profiling in malicious cyber insider investigations. *Digital Investigation*, 3(1), 20–31. doi:10.1016/j.diin.2006.01.006

Steel, C. M. (2014). Idiographic Digital Profiling: Behavioral Analysis Based On Digital Forensics. *The Journal of Digital Forensics. Security and Law: JDFSL*, 9(1), 7–18.

Turvey, B. E. (2011). *Criminal profiling: An introduction to behavioral evidence analysis*. Academic Press.

Yu, S. (2011). *The art of criminal investigation: 50 principles*. CreateSpace Inc.

Yu, S. (2013). Behavioral evidence analysis on Facebook: A test of cyber-profiling. *Defendology*, 16(33), 17–30. doi:10.5570/dfnd.en.1333.02

Zhang, D., Guo, B., Li, B., & Yu, Z. (2010). Extracting social and community intelligence from digital footprints: an emerging research area. In *International Conference on Ubiquitous Intelligence and Computing* (pp. 4–18). Springer. 10.1007/978-3-642-16355-5_4

KEY TERMS AND DEFINITIONS

Computer Forensics: Searching and analyzing digital evidence that can be used to prove criminality.

Cyber Context: The online settings and interactions in which online behavior occurs.

Cyber Profiling: An idiographic analysis on a subject's digital footprints so as to reveal information that can help identify or better understand the subject.

Digital Footprints: The information about a person that exists on various electronic platforms as a result of this person's online behavior or online activity on those platforms.

False Negative: Wrongfully deem digital footprints as irrelevant and therefore exclude them from analysis.

False Positive: Wrongfully attribute digital footprints to a person when this person did not contribute to such digital footprints.

Online Behavior: A person's online activity that results in digital footprints.

Victimology: The study of the victim's role in a crime based on the assumption that the victim's characteristics or actions may have indirectly facilitated or directly resulted in victimization.

Boko Haram's Feminization, Minorization, and Cyberization of Terrorism: Offering the Cyberterrorism Diffusion Model as Anti-Bokoharamism Tool

Adamkolo Mohammed Ibrahim

 <https://orcid.org/0000-0003-1662-7054>

University of Maiduguri, Nigeria

INTRODUCTION

Now, it is no longer any news that Boko haram is dreaded because of its brute acts of terrorism — killing everybody indiscreetly, regardless of faith, gender, age or social status despite the group's claims that it's activities are for Islamic proselytization, a claim that most orthodox Muslims the world over have unanimously condemned. However, over the last four years, whatever killing the group perpetrates is perceived by the public as comparatively less dreadful as a boom of a successfully perpetrated suicide attack. This is because following the weakening over the last couple of years (2016 to 2018) of the terrorists by Nigerian joint military troops, they resorted to using a subtler guerrilla war tactic: laying ambush to Nigerian troops convoys (mostly in remote areas), overrunning and pillaging military camps, bases and even barracks as well as attacking and sometimes laying ambush to motorcades of humanitarian aid personnel zigzagging vulnerable communities in the area. But this subtle war tactic seldom yields large number of casualties as the group would want (given the nature of the targets) compared to suicide bomb attacks which often occur in crowded public places like markets and mosques; hence, people tend to be scared more by suicide attacks, and hence the terrorists' infatuation for girl captives and their preference for suicide bombing as a tool for terrorism (Elden, 2014; GTI, 2017, 2016; UNICEF, 2016a).

So far, the nine-year insurgency has had a devastating effect on the society and people — dozens of thousands have been killed, more than a million others have been forced from their homes and properties worth billions of Dollars have been destroyed (*The Independent*, 2017). Scholars and experts have viewed the attack methodologies adopted by the militants from various perspectives with a view to analysing them and proffering solutions to the protracted insurgency for peace to be restored in the society. Whenever war against terrorism is discussed, the issue of 'sustainable peace' cannot be overlooked. Societies may survive relative peaceful atmosphere but effective social, economic, political and cultural development can only be achieved when there is sustainable peace (Abdulazeez, 2017; Ibrahim & Hassan, 2017).

Currently, the dominant paradigm being employed by the Government in the fight against terrorism has largely been militarized action. Despite the flaws observed in this approach especially human rights abuse by the military, a tremendous success has been achieved within the last couple of years, from 2016 to 2017. The employment of military option to combat insurgency and terrorism may be able to 'enforce' peace, or restore peace through gun nozzle. But an 'enforced' peace (as is being done by Nigerian troops fighting Boko haram) may be difficult to last long (Abdulazeez, 2017). Hence, the need for a multifaceted approach to combat terrorism in Nigeria and worldwide. Historically, over the

DOI: 10.4018/978-1-7998-3479-3.ch025

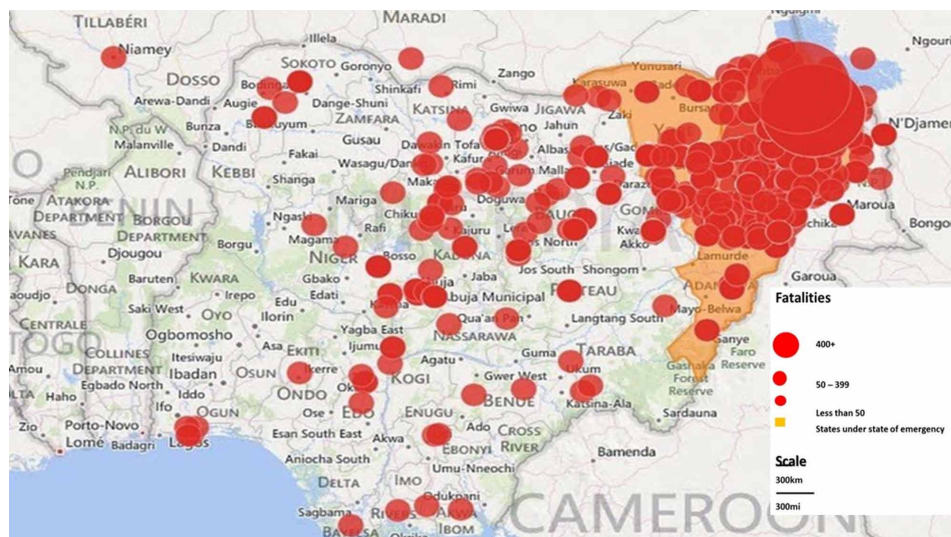
last decade, terrorists have been striving to have a digital footprint in cyberspace (GTI, 2017; Khayat, 2015), this chapter suggests a research model to address issues surrounding cyber terrorism and conflict.

BACKGROUND

A study carried out in the United States (US) has shown that Boko haram's gendered and youth-centric terrorism ballooned after the group understood what *CNN* (2017) reports, "the potency that gender and youth offer in raising its global profile," after the 14 April 2014 kidnapping of the 276 Chibok school-girls, an incident that prompted the 'Bring Back Our Girls' campaign globally. Boko haram has not only realized that leaning against the screen of gender and youth-centrism by using children (mostly girls) as suicide bombers can raise its global profile but also obviously resolved to make this innovative-tactic a tradition and lead other terror groups around the world in the use of children and women as suicide bombers, a phenomenon that, according to Warner and Matfess (2017 August), has shattered demographic stereotypes as to what a suicide bomber looks like.

Since 2009, when it started its terrorism in north-eastern Nigeria (see Figure 1), Boko haram has notoriously achieved two ironic attributes: (i) being the deadliest terror group in the world (even outstripping ISIL and Al-Shabab) from 2014-2015, and (ii) being the first terror group in human history to pioneer the use of more female suicide bombers than male. Available records have shown that Boko haram, a terrorist group that pledged allegiance to ISIL in March 2015 started using suicide bombers on April 8, 2011, that is, about two years after it began insurgency at Suleja town in the Nigerian state of Niger. The attack claimed nine lives including those of eight National Youth Service Corps (NYSC) members (Warner & Matfess, 2017 August). The group started to use women suicide bombers on a large scale in 2014, two months after the Chibok schoolgirls kidnapping. Since then, no other terrorist group in history has used as many female suicide bombers as Boko haram. This incident, "which proved to be the beginning of a terrible end" (Meservey, 2018, paragraph 3, n.p.) was first broke to the world in October 2015 by a global geopolitical intelligence outfit, Stratfor.

Figure 1. A map showing areas of Boko haram attacks in Nigeria and neighboring Cameroon, Chad and Niger, with concentration of the attacks in the country's northeast (Los Angeles Time, 2014)



Prior to 2014, however, the group used only male suicide attackers. (*Financial*, 2017; *The Independent*, 2017; UNICEF, 2016a; UNICEF, 2016b). Again, on February 19, 2018, the militants abducted 110 schoolgirls at Government Girls Science and Technical College in the Yobe State town of Dapchi. In collaboration with Swiss Government and the International Committee for the Red Cross, the Nigerian Government engaged the militants in series of dialogue and negotiations for the safe release of the poor schoolgirls. On March 21, 2018, 30 days after their abduction 105 of them were returned except for a 15-year-old Leah Sharibu, who till the time of this writing has been retained by the terrorists because she refused to denounce her Christian faith (*Channels Tv*, 2018; *News 24*, 2018).

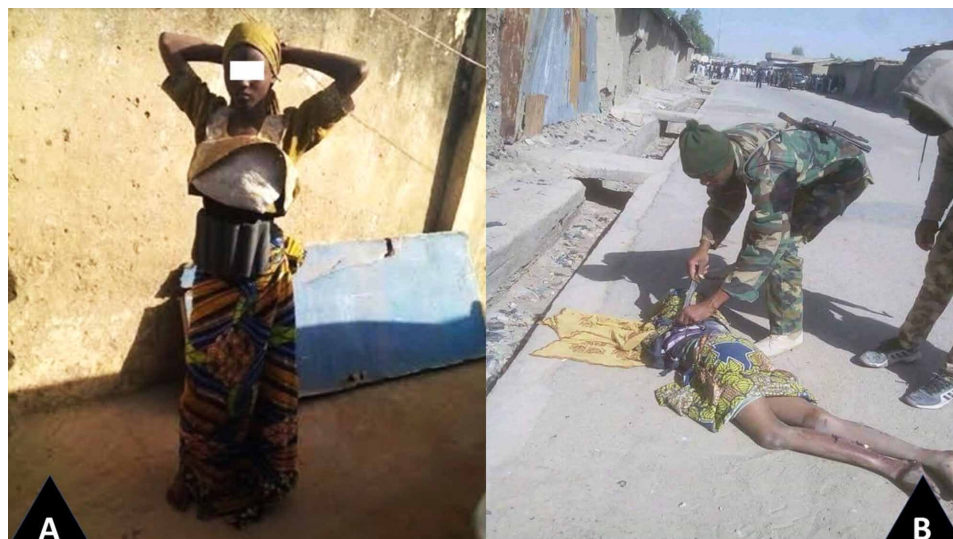
BOKOHARAMISM

Feminization and Minorization of Terrorism

In recent years, there have been regular reports that children are targeted and exposed to attacks and ruthless violence in their homes, street playgrounds and schools. A recent report indicates that from June 2014 to the end of February 2018, Boko haram deployed 469 female suicide bombers in 240 attacks killing 1,259 people (excluding the bombers) and injuring 2,967 others. Most of those incidents (179, 75%) occurred in Nigeria; 48 incidents (20%) occurred in the neighboring Cameroon's region of Far-North; while Chad and Niger republics had six (2%) and seven (3%) respectively. Additionally, majority (133, 55%) of the attacks in Nigeria occurred in Borno State (Pearson, 2018 May).

Additionally, five months after Pearson's (2018 May) report not less than 15 suicide attacks were carried out by the group about 10 of the incidents perpetrated by female 'human bombs' (see Figure 2),

Figure 2. Picture A shows a young female Boko haram suicide bomber rescued by Nigerian security forces before she detonated the explosives strapped around her waist. Picture B shows a young female Boko haram suicide bomber being stripped off the vest of explosives strapped around her waist by Nigerian troops after an unsuccessful suicide mission which culminated in her killing after she refused to obey orders to cast off the IED (Premium Times Nigeria, 2018)



thus yielding a new total of 486 female suicide bombers. The remaining five incidents were perpetrated by young boys. Scores of people were killed, with the highest figure of fatalities in the most recent incidents coming from the twin suicide attacks carried out by two adolescent boys in a market in the Adamawa town of Mubi on May 1, 2018 killing at least 86 and injuring 58 others. In November of the previous year, another attack had killed 50 persons in the same market (*Aljazeera*, 2018 May).

Despite the incidents have drastically reduced since 2016 due to the Government's commitment in the fight against the militants and drawing from Global Terrorism Index's (GTI) estimation that girl more likely to be deployed as suicide attackers than boys at an approximated ratio of four-to-one (GTI, 2017, 2016), one can confidently say that three in every five suicide bombers would be a child and nine in every 10 suicide bombers would be girls. Obviously, so long as Boko haram militancy lasts, more suicide attacks by young girls might be expected. Another key reason the group prefers female suicide attackers is the ease at which females evade security scrutiny than a man strapped with explosive devices could (Sawicki, 2016, July-August).

Bokoharamism is the planned, deliberate and tactical exploitation by terrorist groups of the social value and vulnerability of women and children for deployment to perpetrate suicide bomb attacks to raise the groups' global notoriety, create an unprecedented degree of fear in society and draw global attention, a strategy pioneered by Nigeria's Islamist group, Boko haram. Four years after the emergence of this method of massacre which began in 2014, it continues till the time of this writing. One critical question that needs to be answered has been: 'is Boko haram's deployment of girls and boys suicide attackers a tactic, strategy or concept?'

Firstly, Bokoharamism is a tactic because it involves disposing the children deployed on suicide missions in order of battle-like organizing operations (Sawicki, 2016, July-August), especially the way they are deployed from their base to the manner they approach their targets when they are about to detonate the IEDs strapped on their bodies. This is particularly evident in the group's earlier suicide attack missions. Secondly, Bokoharamism is a strategy because it involves the group's deliberate use of children and women as suicide bombers as a framework or platform upon which the planning and directing a large percentage of its operations and movements within the war theatre (zone) are pivoted with the aim of sustaining its activities and perpetuating its effect in the society (see Warner & Matfess, 2017 August). This is quite evident in the group's current manoeuvres and operations, especially from 2015 when the group's activities turned out to be deadliest till the time of writing this chapter. Thirdly, the term Bokoharamism is a concept since the term itself was constructed from an idea, a thought, a symbol with theoretical underpinning (Elden, 2014; Zenn, 2015 February).

SOLUTIONS AND RECOMMENDATIONS

Cyberization of Terrorism

The literature has shown that Boko haram's global notoriety has been directly or indirectly promoted by cyber technology (Antwi-Boateng, 2017; Elden, 2014; Zenn, 2015 February; Warner & Matfess, 2017 August). Directly, because the group relies on the media to spread propaganda (aimed at waging a psychological warfare on the society) via You Tube and online news sites such as *Sahara Reporters* and *Aljazeera* (Antwi-Boateng, 2017; Elden, 2014; Zenn, 2015 February). Indirectly, because the group benefits from media coverage and reportage of incidents related to its activities. For example, some media outlets report the incidents as 'breaking news' while others as publish them as 'frontpage

headlines.' Giant Tv networks such as *BBC* and *CNN* even go to the extent of attempting to broadcast the incidents live using amateur video footages from citizen journalists (Semetko & Tworzecki, 2017). In fact, with the profound enhancements in Internet technologies, terrorists are making the hay while the sun shines on cyberspace, a phenomenon that calls for serious actions globally to contain their grip on the budding digital media.

The Cyber Terrorism Diffusion Model

The literature shows that terrorists often act as groups, or networks (Rapoport, 2006). This chapter presents a suggested Cyber Terrorism Diffusion Model (CTDM) which focuses on groups' or networks' use of cyber technology (for details see Ibrahim & Hassan, 2017). Generally, there are no standard set of factors for radicalization, extremism or terrorism (Borum, 2017, 2011). However, an extensive literature review has identified six widely cited conceptual models of radicalization as a series of stages (Beadle, 2017). Drawing mainly from the works of Beadle (2017), Borum (2017, 2011), Moghaddam (2005) and Sageman (2008), this model categorizes cyber terrorism into five dimensions, namely (i) instigating factors, (ii) identity factors, (iii) publicity and exposure factors, (iv) group and cyberspace social network factors, and (v) motivational factors. The constructs and variables of the model were derived from those dimensions.

The model has 14 key predictors of use behavior (UB) (or independent variables, with UB as the main dependent variable, DV), three mediator variables and five moderator (control) variables (see Figure 3 and Table 1). The predictors or independent variables (IVs) of the proposed CTDM are perceived attention (PA), perceived recognition (PRC), perceived respectability (PRS), perceived legitimacy (PL), inherent privacy (IP), inherent user-controllability (IUC), perceived self-aggrandizement (PSA), perceived pervasiveness, inherent globality, inherent autonomy, inherent interactivity, crypto-violence agency, crypto-propaganda agency, and cost value.

Figure 3. The proposed Cyber Terrorism Diffusion Model (CTDM)

Note: The original version of this model is presented in Ibrahim and Hassan (2017, p. 50)

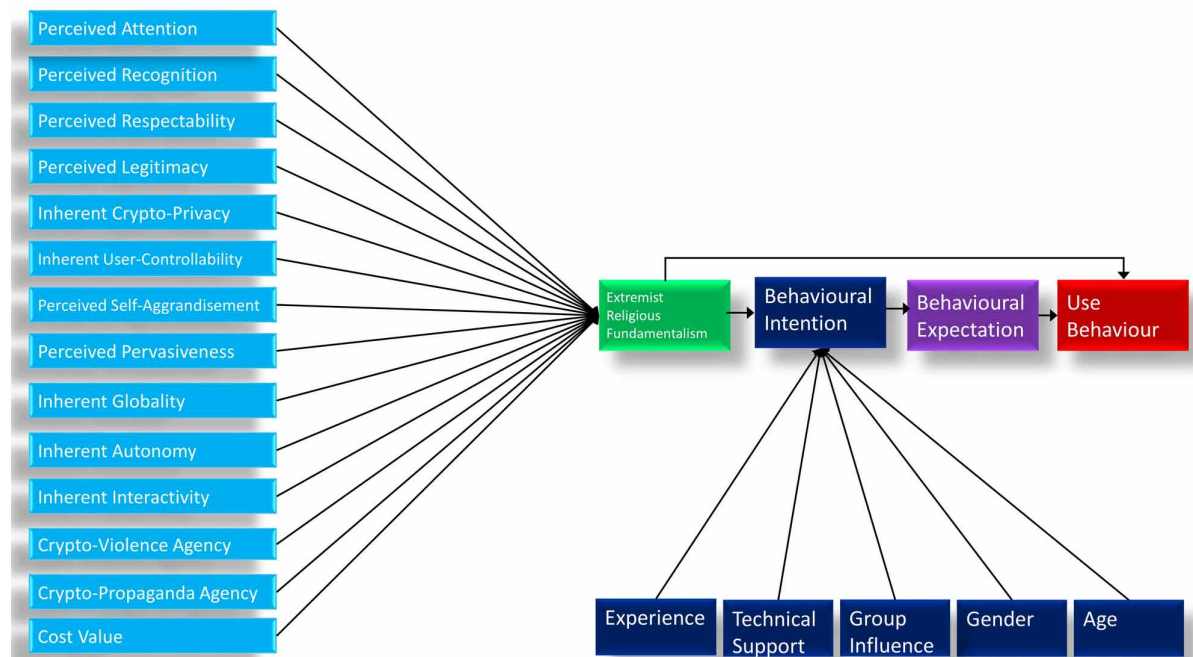


Table 1. Operationalization of the proposed CTDM constructs

S/ No.	Construct/ Variable	Category	Definition	Source
1	PA	IV	The degree to which a group or network of users believe that using a cyber technology can draw wider public attention toward them and render the state attentive to their demands.	Altheide (2009); Beadle (2017); Borum (2017, 2011); Corner and Gill (2015)
2	PRC	"	The degree to which a group or network of users believe that using a cyber technology can render the public and state to be aware, and always be conscious of them.	Beadle (2017); Borum (2017, 2011); Chermask and Freilich (2013)
3	PRS	"	The degree to which a group or network of users believe that using a cyber technology can make them garner wider public and state reverence and regard.	Beadle (2017); Borum (2017, 2011); Corner and Gill (2015)
4	PL	"	The degree to which a group or network of users believe that using a cyber technology to diffuse acts of conflict and terrorism is genuinely justifiable and reasonable.	Altheide (2009); Beadle (2017); Borum (2017, 2011); Chermask and Freilich (2013)
5	ICP	"	The motivation a group or network of cyber technology users derive from the intrinsic or built-in secrecy features of the technology to diffuse acts of conflict and terrorism via the Internet unidentified.	Agre and Rotenberg (1997); Beadle (2017); Barnard-Wills and Ashenden (2012); Byford (1998)
6	IUC	"	The motivation a group or network of cyber technology users derive from the intrinsic or built-in user content-generation, profile customization and general user-friendly applications manipulation features.	Beadle (2017); Borum (2017); Bratich (2011); Carpentier (2011); Chen, Thoms and Fu (2008); Chermask
7	PSA	"	The degree to which a group or network of users believe that using a cyber technology increases their self-worth and dignifies their stature.	Borum (2017, 2011); Chermask and Freilich (2013); Corner and Gill (2015); Gandy (2003)
8	PP	"	The degree to which a group or network of users believe that using a cyber technology to diffuse acts of conflict and terrorism can spread their influence extensively and prevalently.	Beadle (2017); Borum (2017, 2011); Chermask and Freilich (2013); Corner and Gill (2015); Eedle (2002)
9	IG	"	The motivation a group or network users derive from the in-built or 'natural' global ubiquity characteristics of cyber technology.	Baran (2008); Beadle (2017); Borum (2017, 2011); Chen, Thoms and Fu (2008)
10	IA	"	The motivation a group or network of users derive from the vast intrinsic or fundamental freedom and independence of cyber technology from any form of organized control.	Barnard-Wills and Ashenden (2012); Beadle (2017); Borum (2017, 2011); Chermask and Freilich (2013);
11	II	"	The motivation a group or network of users derive from the intrinsic or fundamental characteristics of cyber technology that offers virtually unlimited degree of networked interactivity at both human-to-computer, computer-to-computer and human-to-computer-to-human contexts.	Barnard-Wills and Ashenden (2012); Beadle (2017); Borum (2017, 2011); Corner and Gill (2015)
12	CVA	"	The motivation a group or network of users derive from the virtual world of cyber technology that guarantees tolerance to violence or freedom to incite violence.	Barnard-Wills and Ashenden (2012); Beadle (2017); Chermask and Freilich (2013); Corner and Gill (2015)
13	CPA	"	The motivation a group or network of users derive from the virtual world of cyber technology that guarantees freedom to disseminate propaganda content.	
14	CV	"	The degree to which a group or network of users believe that the use of cyber technology is cost-effective in terms of expenditure and cost-reimbursing in terms of income generation and attracting sponsorship.	Beadle (2017); Biernatzki (2002); Borum (2017, 2011a, b & c); Chermask and Freilich (2013)
15	ERF	MEDV	The religious or secular beliefs and convictions that drive or form parts of the fundamental ideological concepts and legitimizing force of the group or network's movement.	Hobson, Bonk, and Inzlicht (2017); Hoffman (1995); Rapoport (2006, 1983); Sageman (2008); Speckhard (2008)
16	BI	"	The deliberative willingness to use a cyber technology by a group or network to perpetrate violence or terrorism.	Beadle (2017); Borum (2017); Corner and Gill (2015); Maruping et al. (2017)
17	BE	"	A group or network's self-reported subjective probability of its performing a specified behavior, based on its cognitive appraisal of volitional and non-volitional behavioral determinants.	Maruping et al. (2017); Venkatesh, Brown, Maruping, and Bala (2008); Venkatesh, Maruping, and Brown (2006)
18	UB	DV	The actual use, manipulation, running and/ or handling of a cyber technology by a group or network for violence or terrorism.	Beadle (2017); Chermask and Freilich (2013); Corner and Gill (2015); Davis (1989); Venkatesh et al. (2003)

continued on following page

Table 1. Continued

S/ No.	Construct/ Variable	Category	Definition	Source
19	EX	MODV	A group or network of users' prior cyber technology use experience.	Beadle (2017); Chermask and Freilich (2013); Maruping et al. (2017); Venkatesh et al. (2003)
20	TS	''	The belief that a technical help or infrastructure exists to guarantee the use of cyber technology for terrorism and conflict.	Borum (2017); Chermask and Freilich (2013); Corner and Gill (2015); Maruping et al. (2017); Venkatesh et al. (2003)
21	GI	''	The belief that important other groups or networks expect that another group or network should use cyber technology to diffuse acts of violence and terrorism.	Chermask and Freilich (2013); Corner and Gill (2015); Lakhani and Wolf (2005); Maruping et al. (2017); Venkatesh et al. (2003); Webber and Kruglanski (2017)
22	GDR	''	Sex role of members of a group or network of cyber technology users.	Ascol, S., Ayachi, W., Lovell, A., McGill, K., Pagrabs, E., and Yandell R. (2018 April); Speckhard, A. (2008)
23	AG	''	Oldness or youngness of members of a group or network of cyber technology users.	Hecker, T., Fetz, S., Ainamani, H., and Elbert, T. (2015 October); Nandi, C., Elbert, T., Bambonye, M., Weierstall, R., Reichert, M., and Zeller, A. (2017); Speckhard, A. (2008 March)

Note: For details of this table, see Ibrahim and Hassan (2017, p. 50)

pervasiveness (PP), inherent globality (IG), inherent autonomy (IA), inherent interactivity (II), crypto-violence agency (CVA), crypto-propaganda agency (CPA) and cost value (CV). The mediator variables are extremist religious fundamentalism (ERF), behavioral intention (BI) and behavioral expectation (BE) all theorized to mediate the links between the 14 IVs and the DV. BI is theorized to be moderated by five variables, namely experience (EX), technical support (TS), group influence (GI), gender (GR) and age (AG). This is because the core concepts of EX, TS and GI have roots in habit (spontaneous or automatic actions), and all behaviors are governed by intention (Maruping, Bala, Venkatesh & Brown, 2017) and rituals for faith-based actions and practices (Hobson, Bonk, & Inzlicht, 2017) where the power of intention is weak (Borum, 2017, 2011; Maruping et al., 2017; Rapoport, 2006; Wiktorowicz, 2004). BE has been found to be a stronger predictor of use behavior (UB) than BI (Ibrahim & Hassan, 2017; Maruping et al., 2017).

FUTURE RESEARCH DIRECTIONS

Terrorism is evil. However, from all indications, it is here to stay for good. The world must rise against it for peace to reign in all nooks and crannies of human society. Given that conflict is human, and terrorism is borne out of conflict both phenomena may be inexterminable (Elden, 2014). Nonetheless, they can be mitigated. Through the ages, the world has had more than enough share of terror and is still experiencing it. Historically and practically, the dominant paradigm in combatting terrorism and resolving violent conflicts has largely been military force. Yet, like wild fires, violent conflicts and acts of terrorism are continuing here and there across the globe, with no foreseeable end (Abdulazeez, 2017). This paper wants to argue that, comparatively, little of non-military force options have been adopted and successfully applied on a global scale in anti-terrorism campaign. Therefore, a mixed approach involving, e.g., communication for development (e.g., dialogue and alternative dispute resolution means) and research-based solutions with gun power is recommended as one of the efficacious ways of combatting terrorism and conflicts that future research should focus on.

Furthermore, given the spate of developments and innovations in information and communication technologies (ICTs) such as the Internet, website, social media, computer, etc., a great deal of acts of terrorism are progressively planned, coordinated and executed online nowadays. This phenomenon has rendered the terrorists have a digital footprint in cyberspace and society and enabled them to recruit and radicalize new members (Brachman, 2006; King & Taylor, 2001). Hence, this chapter recommends that future research should focus on this new, emerging field of study that expands into the areas of conflicts in cyberspace, including state-of-state cyber warfare, cyber conflict, cyber terrorism and cyber militias.

CONCLUSION

Sadly, as this chapter concludes, Boko haram's activities are continuing with no predictable end to them; at least not soon as signalled by the group's continual suicide bombings and attacks on civilian soft targets, military camps and humanitarian aid workers. Sadly too, the deployment of girls, boys and sometimes, young women in suicide missions continues, a phenomenon that casts a heavy shadow of uncertainty over the future of these poor souls in the Boko haram-infested areas. Even as the fight against the terrorists continues unabatedly, humanitarian aid workers providing livelihoods support to more than a million internally displaced persons (IDPs) and journalists working round the clock covering the conflict could be the terrorists' targets in-the-waiting, especially the former. The world must stand together to reaffirm its commitment to protect all vulnerable groups in the conflict zone. Humanitarian workers and journalists should never be a target much as children and women should not be. In fact, under no circumstance whatsoever should anyone be a target of the militants' killing, maiming and abduction.

It is long overdue Boko haram laid down its arms and accepted the amnesty corridor offered by the Nigerian Government. Redress to grievances and differences should be sought through civil and legal means and all ways of expressing grievances capable of breeding conflict and crisis must be avoided. The UN should continue to play a key role in the global war against terrorism, particularly in West Africa and Nigeria. Meanwhile, UNICEF and other humanitarian aid organizations should continue to work with the Nigerian Government and those of neighboring countries to ensure the safe return of all the abducted schoolgirls, humanitarian aid workers and, indeed, all kidnapped persons. In conclusion, this chapter summons the Government and all critical stakeholders to employ all the scientific and non-scientific methods of exterminating Boko haram terrorism for good through these three critical ways:

- (i) building, operating and sustaining a strong, systematic and professional intelligence network;
- (ii) adopting communication for development strategies, e.g., dialogue and alternative dispute resolution as potent tools of anti-terrorism; and
- (iii) adopting the CTDM approach suggested in this chapter in scientific research on issues surrounding not only Boko haram insurgency but all forms of conflict with a view to stopping violent conflicts and maintaining the peace in the society.

REFERENCES

- Abdulazeez, M. A. (2017). The inadequacies of exclusive military action: Countering Boko haram's terrorism with alternative approaches. In *Selected Papers: Terrorism Experts Conference (TEC) 2016* (pp. 26–60). NAT.
- Aljazeera. (2018 May). Nigeria mosque attack death toll rises to 86. *Aljazeera*. Retrieved August 1, 2018 from <https://www.aljazeera.com/news/2018/05/nigeria-mosque-attack-death-toll-rises-86-180502162539760.html>
- Antwi-Boateng, O. (2017). Boko haram and the Islamic State: A tale of two terrors. *Conflict Studies Quarterly*, 18, 20–39.
- Beadle, S. (2017, March 19). *How does the Internet facilitate radicalization? Homegrown radicalization and counter-radicalization*. London: War Studies Department, King's College London.
- Bloom, M. & Matfess, H. (2016). Women as symbols and swords in Boko haram's terror. *PRISM, a Journal of the Center for Complex Operations*, 6(1), 105 – 121.
- Borum, R. (2011). Understanding terrorist psychology. In A. Silke (Ed.), *The psychology of counter-terrorism* (pp. 19–33). Routledge.
- Borum, R. (2017). The etiology of radicalization. In G. LaFree & J. D. Freilich (Eds.), *The handbook of the criminology of terrorism* (pp. 17–32). Wiley Blackwell. doi:10.1002/9781118923986.ch1
- Channels, Tv. (2018). Dapchi schoolgirls' abduction a national disaster – Buhari. *Channels Tv*. Retrieved March 22, 2018 from <https://www.channelstv.com/2018/02/23/dapchi-schoolgirls-abduction-national-disaster-buhari>
- CNN. (2017). Boko haram favors women, children as suicide bombers, study reveals. *CNN*. Retrieved January 23, 2018 from <http://edition.cnn.com/2017/08/10/africa/boko-haram-women-children-suicide-bombers/index.html>
- Elden, S. (2014). The geopolitics of Boko haram and Nigeria's 'war on terror'. *The Geographical Journal*, 180(4), 414–425. doi:10.1111/geoj.12120
- GTI. (2016). *Global terrorism index: Measuring and understanding the impact of terrorism*. Institute for Economics and Peace.
- GTI. (2017). *Global terrorism index: Measuring and understanding the impact of terrorism*. Institute for Economics and Peace.
- Hobson, N. M., Bonk, D., & Inzlicht, M. (2017). Rituals decrease the neural response to performance failure. *PeerJ*, 5, e3363. doi:10.7717/peerj.3363 PMID:28584707
- Ibrahim, A. M., & Hassan, M. S. (2017). *Media versus Boko haram: Proposing the Adoption and Use of Cyber Technology by Terror Groups Model for counterinsurgency*. Lambert Academic Publishing.
- Khayat, M. (2015). *Jihadis shift to using secure communication app telegram's channels service*. MEMRI - The Middle East Media Research Institute. Retrieved October 29, 2015 from <https://www.memri.org/>
- Los Angeles Time. (2014). Mapping every Boko haram attack. *Los Angeles Time*. Retrieved August 7, 2017 from <https://www.latimes.com/visuals/graphics/la-fa-g-boko-haram-20140707-htmllstory.html>

- Maruping, L. M., Bala, H., Venkatesh, V., & Brown, S. A. (2017). Going beyond intention: Integrating behavioral expectation into the unified theory of acceptance and use of technology. *Journal of the Association for Information Science and Technology*, 68(3), 623–637. doi:10.1002/asi.23699
- Meservey, J., & Vadyak, A. (2018, June 20). Boko haram's sick ploy to turn girls into suicide bombers. The Heritage Foundation. Retrieved July 30, 2018 from <https://www.heritage.org/terrorism/commentary/boko-harams-sick-ploy-to-turn-girls-suicide-bombers#>
- Moghaddam, F. (2005). The staircase to terrorism. *The American Psychologist*, 60(2), 161–169.
- News 24. (2018). Nigeria's Buhari meets with Dapchi girls. *News 24*. Retrieved March 24, 2018 from <https://m.news24.com/Africa/News/nigerias-buhari-meets-with-dapchi-girls-20180323>
- Pearson, E. (2018, May). *Boko haram beyond the headlines: Analysis of Africa's enduring insurgency*. West Point, NY: Combatting Terrorism Center at West Point.
- Premium Times Nigeria. (2018). Soldiers kill two female suicide bombers, arrest another. *Premium Times*. Retrieved March 26, 2018 from <https://www.premiumtimesng.com/news/headlines/255012-soldiers-kill-two-female-suicide-bombers-arrest-another.html>
- Rapoport, D. C. (Ed.). (2006). *Terrorism: The fourth or religious wave* (Vol. 4). Taylor & Francis.
- Financial. (2017). Female suicide bombers as victims of Boko haram insurgency. *Financial Magazine*. Retrieved December 15, 2017 from <http://www.financialnigeria.com/female-suicide-bombers-as-victims-of-boko-haram-insurgency-blog-302.html>
- Sageman, M. (2008). *Leaderless jihad: Terror networks in the Twenty-First Century*. University of Pennsylvania Press. doi:10.9783/9780812206784
- Sawicki, J. (2016, July-August). A tragic trend: Why terrorists use female and child suicide bombers. *Health Progress (Saint Louis, Mo.)*, 97(4), 38–42. PMID:28165685
- Semetko, H. A., & Tworzecki, H. (2017). *Content effects: Political communication*. *The International Encyclopedia of Media Effects*. John Wiley & Sons, Inc. doi:10.1002/9781118783764.wbieme0124
- The Independent. (2017). Boko haram: Huge rise in children used as “human bombs” by Islamists in Nigeria, UN warns. *The Independent*. Retrieved November 3, 2017 from <http://www.independent.co.uk/news/world/africa/boko-haram-child-suicide-bombers-west-africa-unicef-nigeria-islamist-terrorists-a7907166.html?amp>
- UNICEF. (2016a). *Beyond Chibok: Over 1.3 million children uprooted by Boko haram violence*. Retrieved January 25, 2018 from http://www.unicef.org/infobycountry/files/Beyond_Chibok.pdf
- UNICEF. (2016b). Nigeria-humanitarian SITREP No. 17. *Nigeria country office situation report*. UNICEF. Retrieved December 28, 2017 from http://www.unicef.org/NG_wwwd_sitreps17.pdf
- Warner, J. & Matfess, H. (2017 August). *Exploding stereotypes: The unexpected operational and demographic characteristics of Boko haram's suicide bombers*. Washington, DC: Combating Terrorism Center at West Point.
- Wiktorowicz, Q. (2004). *Islamic activism: A social movement theory approach*. *Indiana Series in Middle East Studies*. Indiana University Press.

Zenn, J. (2015, February). *The power of Boko haram: Local roots, international jihad*. London: Tony Blair Faith Foundation. Retrieved August 24, 2016 from <http://tonyblairfaithfoundation.org/religiongeo-politics/commentaries/opinion/powerboko-haram-local-roots-international-jihad>

ADDITIONAL READING

Agara, T. (2015). Gendering terrorism: Women, gender, terrorism and suicide bombers. *International Journal of Humanities and Social Science*, 5(6/1), 115 – 125.

Crelinsten, R. (2017). A comprehensive approach to counterterrorism: An overview. In *Selected papers: Terrorism Experts Conference (TEC) 2016* (pp. 8–25). NATO.

Dawson, M., & Adeboje, W. (2017). Islamic extremists in Africa: Security spotlight on Kenya and Nigeria. In M. Dawson, D. R. Kisku, P. Gupta, J. K. Sing, & W. Li, (Eds.), *Developing next-generation countermeasures for homeland security threat prevention* (pp. 93 – 103). Hershey, P.: IGI Global.

Matfess, H. (2016, April 26). Here's why so many people join Boko haram, despite its notorious violence. *The Washington Post*. Retrieved July 1, 2017 from https://www.washingtonpost.com/news/monkey-cage/wp/2016/04/26/heres-why-so-many-people-join-boko-haram-despite-its-notorious-violence/?utm_term=.bfa1b064d122

Muhammed, K. (2014). The message and methods of Boko haram. In P. M. Marc-Antoine (Ed.), *Boko haram: Islamism, politics, security and the state in Nigeria* (p. 13). African Studies Centre.

Ottis, R., & Lorents, P. (2010, April). Cyberspace: Definition and implications. In *International Conference on Cyber Warfare and Security* (p. 267). Academic Conferences International Limited.

Phillips, P. J., & Pohl, G. (2017). Terrorist choice: A stochastic dominance and prospect theory analysis. *Defence and Peace Economics*, 28(2), 150–164. doi:10.1080/10242694.2015.1033888

Saulat, P. (2010, June 25). 'Why does Islam have the concept of jihad or holy war, which some use to justify violence and terrorism?' *Why Islam* Islamic blog. Retrieved August 15, 2017 from <https://www.whyislam.org/jihad-2/jihad-faqs/islamic-concept-of-jihad-holy-war/>

KEY TERMS AND DEFINITIONS

Abduction/Kidnapping: The act of capturing, seizing or forceful carrying off vulnerable individuals by armed non-state actors for political, economic, religious, or social goals.

Boko Haram: A group of Islamist insurgents in the northeast of Nigeria who rose against the Government since 2009 and metamorphosed into a terrorist organization over the years. The group's activity has spilled over into the neighboring countries of Cameroon, Chad and Niger and has been responsible for the death of nearly 30,000 people and displacement of nearly 1.5 million others in Nigeria alone.

Bokoharamism: The planned, deliberate, and tactical exploitation by a terrorist group of the social value and vulnerability of minors, specifically creating an unprecedented degree of fear in society and drawing global attention. This strategy was pioneered by Nigeria's Islamist group, the Boko haram.

Bomb: An improvised explosive device (IED) used as a tool for terrorism.

Cyberization: The use of the Internet, or cyberspace on a sustainable basis by a terrorist group, militias, or other similar groups engaged in conflicts to promote and propagate their causes.

Feminization: The deployment of female (often girls and young women) captives on a sustainable basis by terrorist groups and networks as ‘human bombs’ in suicide missions.

Human Bomb: The use by terrorist groups of (human) captives as bombs (or explosives) on a sustainable basis. Terrorist groups such as Boko haram, ISIL and Al Qaeda have been notorious, particularly Boko haram for the deployment of their captives (mostly young girls and boys) as suicide bomb attackers strapped with IED jackets.

Minorization: The deployment of vulnerable persons (girls, boys, and women) captives by terrorist groups and networks as suicide bombers on a sustainable basis.

Suicide: The act of killing oneself with others using IEDs voluntarily or by coercion for political, economic, religious, or social goals.

Terrorism: The threatened or actual use of illegal force and violence by a group or network of non-state actors (whose main targets are not their actual targets) to gain political, economic, religious, or social goals by fear, coercion, or intimidation.

Cyberbullying: Definition, Behaviors, Correlates, and Adjustment Problems

Michelle F. Wright

Pennsylvania State University, USA & Masaryk University, USA

INTRODUCTION

Many youths have embraced digital technologies, such as mobile phones and the Internet. They often use these technologies daily (Lenhart, 2015). There are many benefits associated with youths' use of digital technologies, such as having the opportunity to communicate with anyone at anytime and in any location, access to information for leisure and school purposes at the click of a button, and easy access to entertainment (e.g., playing online games, watching videos). Despite these benefits, there are opportunities for negative interactions and experiences. These negative interactions and experiences might involve receiving unwanted electronic content via videos, images, and text, identity theft, misuse of personal information, utilizing fake or untrue information for school assignments, and sexual predation. Cyberbullying is another risk associated with youths' digital technology use.

Cyberbullying is often conceptualized as an extension of traditional bullying (Bauman, Underwood, & Card, 2013a; Grigg, 2012). Similar to traditional bullying, cyberbullying involves bullying behaviors that are perpetrated with malicious intent to cause harm to victim or victims through digital technology, such as email, instant messaging, social networking websites, and text messaging through mobile devices. The cyber context allows youths to keep their identity private or to remain anonymous. Because cyberbullies can remain anonymous, they often can harm their victims while also removing any concern for the consequences of their actions (Wright, 2014b). Furthermore, anonymity might lead youths to do or say things online that they would never do or say in the offline world. This is known as the online disinhibition effect (Suler, 2004; Wright, 2014a). Another characteristic of cyberbullying is that cyberbullies are capable of harming their victims in a shorter period of time. In particular, cyberbullies can spread a rumor much quicker in the online world than in the offline world. It might take minutes for a rumor to spread to the whole classroom, while it could take hours for this to occur in the offline world. In addition, cyberbullies can target victims as many times as they want and almost anywhere as well. No longer are victims safe in their own homes, as cyberbullying can follow them into their homes and other places they believe are safe. Cyberbullying can involve the bully and victim only, a bully and multiple victims, one bystander, or multiple bystanders. For example, a cyberbully might post a video making fun of someone and this video might receive thousands of watches, whereas being bullied in the lunchroom might be only visible to youths playing attention to what is happening. Due to these characteristics, cyberbullying is distinctive from traditional face-to-face bullying.

The purpose of this chapter is to review literature on cyberbullying among youths in elementary, middle, and high schools. The literature review draws on studies from different disciplines, including psychology, education, media studies, communication, social work, sociology, and computer science. These studies might also include those with cross-sectional, longitudinal, qualitative, quantitative, mixed

DOI: 10.4018/978-1-7998-3479-3.ch026

methods research designs. Another purpose is for these studies to explain cyberbullying in different countries in an effort to provide a much thorough review of the literature. The chapter is divided into six sections: a) background of cyberbullying, b) characteristics and risk factors associated with cyberbullying, c) negative consequences of cyberbullying, d) solutions and recommendations, e) future research directions, and f) conclusion.

BACKGROUND

Cyberbullying involves the use of digital technologies to hostilely and intentionally harass, embarrass, and intimidate others (Smith et al., 2013). The hostility and intentionality portions of this definition are key to defining cyberbullying, as these behaviors must include a desire to maliciously harm the victim or victims. Similar to traditional forms of face-to-face bullying, cyberbullying can sometimes include repetition and an imbalance of power. Repetition of cyberbullying acts is complex as sharing humiliating videos or text messages can be sent to the victim only, one bystander, or multiple bystanders (Bauman et al., 2013b). The nature of repetition in the cyber context makes it easier for bullies to continue the cycle of cyberbullying behaviors. It is possible for cyberbullies to distribute a humiliating video or text message to one person, and then for this person to share the content additional times with other people. The people can then also share the video or text message.

The use of digital technologies is a component of the cyberbullying definition that separates it from the definition of traditional face-to-face bullying (Curelaru, Iacob, & Abalasei, 2009). There are various digital technologies that can be used to perpetrate cyberbullying behaviors, such as sending unkind text messages and emails, theft of identity information, pretending to be someone else, making anonymous phone calls, sharing secrets about the victim by posting or sending the secret to someone else, spreading nasty rumors using social networking websites, threatening to harm someone, or uploading an embarrassing picture or video of the victim with malicious intent (Bauman et al., 2013a). Some other examples of cyberbullying behaviors are similar to those that occur in the offline world, such as harassment, insults, verbal attacks, teasing, physical threats, social exclusion, and humiliation. Cyberbullying can involve distributing explicit or embarrassing video through different mediums, including social networking sites, text messages, and online gaming sites, creating websites to defame someone else, and making fake social networking profiles using someone else's identity (Rideout et al., 2005).

Some cyberbullying behaviors have no offline equivalent, such as happy slapping and flaming (Rideout et al., 2005). Happy slapping is defined as a group of people who randomly insult another person while filming the incident on a mobile phone, and then these individuals will post the image(s) or video(s) online for others to see. Flaming involves posting provocative or offensive messages in a public forum with the intention of provoking an angry response or argument from members of the forum. Although cyberbullying behaviors can occur through different digital technologies, the most frequently used technologies include gaming consoles, instant messaging tools, and social networking websites (Ybarra et al., 2007).

Rates of Cyberbullying Behaviors

It is important to understand the prevalence rates of youths' involvement in cyberbullying as these rates indicate the likelihood that they will be involved in these behaviors. In one study on cyberbullying, Patchin and Hinduja (2006) found that 29% of youth in their sample reported that they were victimized by cyberbullying and that 47% indicated that they had witnessed cyberbullying as bystanders. Similarly,

Kowalski and Limber (2007) found that 11% of the 3,767 middle schoolers, ages 11 through 14, in their study were cyberbullied, 4% bullied others, and 7% were both cybervictim and cyberbully. In another study, Goebert and colleagues (2011) found that 56.1% of Hawaiian youths in their sample, those in grades 9th through 12th, reported that they were cybervictims. Assessing cyberbullying within the past 30 days among 6th through 12th graders, Hinduja and Patchin (2012) found that 4.9% of the youths in their sample admitted to perpetrating cyberbullying.

Cyberbullying studies have also been conducted in Canada. Cappadocia and colleagues (2013) found that 2.1% of Canadian adolescents in their study admitted that they had perpetrated cyberbullying, 1.9% admitted that they were victims of cyberbullying, and 0.6% were classified as both victims and bullies. In another study published in the same year, Bonanno and Hymel (2013) revealed that 6% of adolescents in their sample were cyberbullies, 5.8% were cybervictims only, and 5% were both cyberbullies and cybervictims.

Rates of cyberbullying involvement have been examined among youths in Africa, Asia, Australia, and Europe. Ultimately, these studies highlight the need to focus on cyberbullying. In one study conducted in Sweden, Laftman and colleagues (2013) found that 5% of their 22,544 sample, ages 15-18, were cybervictims, 4% perpetrated cyberbullying, and 2% were classified as both cybervictims and cyberbullies. With a slightly younger sample of Swedish adolescents, those in the 7th through 9th grade, Beckman et al. (2012) classified 1.9% of their sample as cybervictims, 2.9% as cyberbullies, and 0.6% as both cyberbullies and cybervictims. Rates of cyberbullying involvement were similar to those found among adolescents in Ireland. Corcoran and colleagues (2012) found that 6% of adolescents in their sample reported that they were victims of cyberbullying. Slightly higher rates were found by Brighi and colleagues (2012) among their Italian adolescent sample. In this study, 12.5% were identified as cybervictims. Festl and colleagues (2013) found similar prevalence rates among their sample of German adolescents such that 13% were classified as cyberbullies and 11% as cybervictims. Rates of cyberbullying involvement are slightly higher in Israel than other countries in Europe. In particular, Olenik-Shemesh and colleagues (2010) found that 16.5% of the 242 adolescents in their sample were cybervictims. In Lazuras et al.'s (2013) study, they found that 32.4% of their Israeli adolescent sample reported that they were bystanders of cyberbullying.

Many studies have been conducted on the prevalence rates of cyberbullying among Turkish youths. Estimates of cyberbullying victimization among Turkish youths vary, from 18% to 32% (Eruder-Baker, 2010; Yilmaz, 2011). Cyberbullying perpetration rates ranges from 6% through 9% among Turkish youths in a few studies (e.g., Ayas & Horzum, 2012; Yilmaz, 2011). Rates of cyberbullying perpetration were higher in Aricak et al.'s (2008) study. They found that 36% of Turkish secondary school students in their study admitted to being cyberbullies.

Although some studies of cyberbullying involvement have been conducted in Asian countries, this research has been slower to develop. Huang and Chou (2010) reported that 63.4% of the Taiwanese youths in their sample were cyberbystanders, 34.9% were cybervictims, and 20.4% were cyberbullies. Jang and colleagues (2014) found that 43% of the 3,238 Korean adolescents in their study were either cyberbullies or cybervictims. Rates of cyberbullying involvement are similar in studies conducted with youths in China. In their study, Zhou et al. (2013) found that of the 1,438 youths in their sample that 34.8% were classified as cyberbullies and 56.9% as cybervictims. With a focus exclusively on Facebook cyberbullying, 59.4% of Singaporean adolescents in Kwan and Skoric's (2013) study indicated that they were victimized through Facebook and 56.9% admitted to perpetrating cyberbullying. Little attention has been given to understanding cyberbullying involvement among youths in Africa, India, and South America.

Cross-Cultural Differences in Cyberbullying Involvement

Some attention has been given to comparing the prevalence rates of cyberbullying involvement across different countries. Most of this cross-cultural research has compared countries based on independent self-construal (e.g., view self as separate from the social context) or an interdependent self-construal (e.g., view self within the context of their social environment or society). Typically people from Western countries, like the United States, Canada, and England, are reinforced and primed for behaving in ways aligned with an independent self-construal, whereas people from Eastern countries, like China, Korea, and Japan, are reinforced and primed for behavior consistent with an interdependent self-construal. Differences in these self-construals affect people's social behaviors, particularly their involvement in bullying and cyberbullying. Therefore, independent and interdependent self-construals have been used to explain these behaviors.

In one cross-cultural study on differences in cyberbullying, Barlett et al. (2013) found that youths from the United States engaged in more cyberbullying than youth from Japan. Strohmeier and colleagues (2013) found similar results in their study, with Austrian adolescents reporting more cyberbullying perpetration and victimization than Japanese adolescents. Furthermore, Li (2008) reported that Chinese youths engaged in less cyberbullying perpetration in comparison to Canadian youths, although she found no differences in cyber victimization among these samples. Using a different sample, Li (2006) found support similar results, except that Chinese youths reported that they were victimized online more than Canadian youths. Classifying cyberbullying into reactive (i.e., response to provocation) and proactive (i.e., to obtain some sort of goal), Shapka and Law (2013) found that Canadian youths reported less proactive cyberbullying perpetration when compared to East Asian youths. On the other hand, Canadian youths reported more reactive cyberbullying perpetration than East Asian youths.

In the one of the few studies to focus on cross-cultural differences among Asian countries, Wright et al. (2015) found that Indian adolescents reported more cyberbullying perpetration and victimization when compared to adolescents from China and Japan, with Chinese adolescents reporting more of these behaviors in comparison to Japanese adolescents. Wright and colleagues also found that boys from India were more often involved in cyberbullying than boys from China or Japan.

Variations in prevalence rates are due to differences in samples, sampling techniques, and measures used. Considering the literature on cyberbullying involvement in different countries, it is clear that these behaviors are a global concern, warranting additional investigations.

CHARACTERISTICS OF CYBERBULLYING INVOLVEMENT

The first investigations of cyberbullying focused on prevalence rates. However, the differences found across various samples lead to researchers focusing their attention on the characteristics that are associated with youths' involvement in cyberbullying. One of the first variables examined was age. There are mixed findings regarding age differences in cyberbullying. For example, Williams and Guerra (2007) found that physical forms of cyberbullying, like hacking, peaked in middle school, while rates of physical forms of cyberbullying decreased in high school. Opposite findings from these results were found by Wade and Beran (2011), as they found that 9th graders in their sample had the highest risk of being involved in cyberbullying when compared to adolescents in middle school. Consequently, age was not found to be a reliable predictor of youths' cyberbullying involvement.

Gender has also been examined as a risk factor of cyberbullying. Some researchers have found that boys reported that they were more often the perpetrators of cyberbullying when compared to girls (Boulton et al., 2012; Li, 2007; Ybarra et al., 2007). Other research has found that girls were more often victimized by cyberbullying when compared to boys (Hinduja & Patchin, 2007; Kowalski & Limber, 2007). Yet, some researchers (e.g., Dehue, Bolman, & Vollink, 2008; Pornari & Wood, 2010) reported that girls engaged in more cyberbullying perpetration, and that boys experienced more cyber victimization than girls (e.g., Huang & Chou, 2010; Sjurso, Fandrem, & Roland, 2016). There are also other researchers (e.g., Stoll & Block, 2015; Wright & Li, 2013b) who have found no gender differences in youths' involvement in cyberbullying. Given these mixed findings, some researchers have argued for the need to better understand how gender might impact the mediums used and behaviors perpetrated in cyberbullying incidents. In one study on this topic, Wright (in press) found that boys were more likely to be victimized by cyberbullying through gaming consoles and girls perpetrated relational and verbal forms of cyberbullying. Similar to age, gender is an inconsistent predictor of cyberbullying.

Previous involvement in offline forms of bullying are examined as a risk factor associated with cyberbullying involvement. In this research, relationships are usually found between cyberbullying perpetration and traditional face-to-face bullying perpetration, cyber victimization and traditional face-to-face victimization, and traditional face-to-face victimization and cyberbullying perpetration (Barlett & Gentile, 2012; Mitchell et al., 2007; Wright & Li, 2013a; Wright & Li, 2013b).

Another characteristic examined in association with cyberbullying involvement among youths is digital technology. When youths utilized the internet more often, they are also more likely to report being involved in cyberbullying (Ang, 2016; Aricak et al., 2008). Furthermore, cyber victims self-reported using instant messaging tools, email, blogging sites, and online game more often than nonvictims, another finding suggesting the role of digital technology use in cyberbullying involvement (Smith et al., 2008). The link between digital technology use and cyberbullying involvement might be explained by some youths' greater likelihood to disclose personal information online. Disclosing personal information online, like one's geographical location or school name, increases youths' risk for cyberbullying victimization.

Internalizing difficulties, like depression, anxiety, and loneliness, and externalizing difficulties, such as alcohol and drug use, are also related to youths' involvement in cyberbullying. Researchers propose that internalizing and externalizing difficulties reduce youths' ability to utilize effective coping strategies, making them more vulnerable to cyberbullying (Cappadocia et al., 2013; Mitchell et al., 2007). Alcohol and drug use are both associated with youths' involvement in cyberbullying as the perpetrator and victim (Cappadocia et al., 2013; Wright, in press).

Other variables also influence youths' cyberbullying involvement. Higher normative beliefs (i.e., beliefs about the acceptability of behavior) regarding face-to-face bullying and cyberbullying were related positively to cyberbullying perpetration (e.g., Burton, Florell, & Wygant, 2013; Wright, 2014b). Cyberbullies often hold favorable attitudes toward the perpetration of bullying behaviors, which thereby influences them to believe that these behaviors are acceptable to perpetrate. Holding provictim attitudes (i.e., belief that bullying is unacceptable and that defending victims is valuable) is another variable that impacts cyberbullying involvement. Holding lower levels of provictim attitudes, and having lower peer attachment, less self-control and empathy, and greater moral disengagement were each correlates positively with cyberbullying perpetration (e.g., Sevcikova, Machackova, Wright, Dedkova, & Cerna, 2015; Wright et al., 2015).

Many of the studies on the characteristics that influence youths' cyberbullying involvement do not utilize longitudinal designs. In a longitudinal study, Fanti and colleagues (2012) examined children's and adolescents' exposure to violent media, their callous and unemotional traits, and their cyberbul-

lying involvement one year later. Media violence exposure was positively linked to adolescents' cyber victimization one year later. In another longitudinal study, Wright (2014a) found that perceived stress from parents, peers, and academics/school were related to adolescents' cyberbullying perpetration one year later. The research in this section reveals various risk factors and characteristics that increase youths' vulnerability to cyberbullying perpetration and victimization. Most of the literature focuses on *individual-level* risk factors and characteristics that increase youths' vulnerability to cyberbullying. The next sections focus on the role of parents, schools/teachers, and peers in youths' involvement in cyberbullying. Parents, schools/teachers, and peers represented *contextual-level* risk factors.

Parents

Similar to the offline world, some parents monitor their children's activities in the cyber context. In one study, 50% of youths in one sample reported that their parents did not monitor their online activities, although 80% of parents indicated that they monitored their children's online activities (Mason, 2008). In another study, McQuade, Colt, and Meyer (2009) found that 93% of parents indicated that they set limits on their children's online activities, despite 37% of their children reporting that their parents had given them rules concerning their online activities. Both of these studies might indicate that parents overreport how much monitoring they engage in, that parents do not share their monitoring strategies with their children, or that their strategies for monitoring their children's online activities is ineffective because their children do not believe any strategies have been implemented. Parents' role in protecting their children from online risks is important. More research attention should be given to how parents navigate conversations with their children about online risks and opportunities.

Wright (2015) found that parents have a significant role in protecting their children from cyber victimization. Parents who engaged in more mediation of their children's digital technology use reduced the risk of cyberbullying victimization and depression. She concluded that parental mediation buffers against the negative consequences associated with cyberbullying. It could be possible that parents who monitor their children's digital technology use might have more opportunities to discuss the risks associated with technology use. These discussions might increase youths' awareness of the risks associated with cyberbullying and convey to them that negative online behaviors are not acceptable and are punishable. Such a proposal is consistent with research revealing that youths who believe that their parents will punish them for engaging in cyberbullying and other negative behaviors were less likely to engage in cyberbullying (Hinduja & Patchin, 2013; Wright, 2013a). Some research has found that parental monitoring did not influence youths' exposure to online risks. Aoyama and colleagues (2011) found that parental mediation had no impact on youths' involvement in cyberbullying. Their explanation was that many parents lack technological skills and this makes it difficult for them to be sure of how and when to intervene. Such a proposal is consistent with previous research findings in which parents reported that they were not sure how to discuss online activities with their children (Rosen, 2007). This uncertainty might lead some parents to not know to talk to their children about appropriate online behaviors. However, this research is dated and many parents are enjoying the benefits of technology and using digital devices.

Parents who do not enforce or follow-up on the strategies they implement to make the Internet safer for their children might be increasing their children's exposure to online risks. Their children could develop the perception that their parents are unconcerned with their online activities, subsequently increasing their risk of perpetrating cyberbullying. Not only is it important for parents to enforce and implement strategies concerning their children's digital technology use but it is also important for parents to update these strategies as children become older and desire more privacy and independence.

Some research has focused on the family characteristics that could potentially increase youths' risk of cyberbullying involvement. Family income, parental education, and marital status of parents/caregivers was unrelated to youths' cyberbullying involvement (Ybarra & Mitchell, 2004). In later research, Arslan and colleagues (2012) found that parental unemployment was related to youths' cyberbullying perpetration and victimization. Neglectful parenting increased youths' risk of cyberbullying involvement when compared to uninvolved youths (Dehue, Bolman, Vollink, & Pouwelse, 2012). Furthermore, authoritarian parenting style increased youths' experience of cyberbullying as cybervictims.

The literature discussed in this section indicates that parents have an important function in their children's digital lives. The next sections examine literature on the role of schools/teachers and peers in youths' cyberbullying involvement.

Schools

There is much debate concerning schools' role in monitoring and enforcing sanctions for youths' cyberbullying. One of the reasons for this debate is because many cyberbullying incidents occur off school grounds, often making it difficult for schools to know when these situations occur and how to deal with such situations (deLara, 2012; Mason, 2008). Regardless of where cyberbullying incidents occur, many of these incidents involve youths who attend the same school, which could further complicate the schools' role in handling cyberbullying. This is because youths who know each other and who also engage in cyberbullying might increase the likelihood that cyberbullying or knowledge of a cyberbullying situation could spill over into the school environment. This spillover effect could disrupt the learning process and make victims even more fearful.

Despite the potential for cyberbullying to spill over onto school grounds, teachers and administrators can often have very different perceptions of cyberbullying as some of these individuals might perceive cyberbullying as more problematic than others (Kochenderfer-Ladd & Pelletier, 2008). Many times administrators and teachers do not perceive cyberbullying or other types of covert bullying behavior, such as relational bullying, as serious and harmful behaviors (Sahin, 2010). Given these differences, it is not surprising that some teachers tend to encourage prevention programs designed to reduce physical forms of face-to-face bullying more often than other forms of bullying (Tangen & Campbell, 2010). These teachers do not usually understand that there are also harmful consequences associated with bullying and cyberbullying.

There is teacher training that can help teachers recognize cyberbullying and how to deal with it, though this training is often insufficient and does not properly inform teachers of what to do when one of their students is cyberbullied. Teachers can also be unfamiliar with newer technologies, another barrier to them dealing effectively with cyberbullying (Cassidy et al., 2012a). They might not know how to respond to the incident or what strategies to use to alleviate the situation. Even when teachers are concerned about cyberbullying, there are few policies and programs in place at the school level, making it harder for them to implement solutions to help victims (Cassidy, Brown, & Jackson, 2012b).

There are some schools that recognize the importance of implementing effective policies and professional development designed to help teachers, staff, and administrators deal effectively with cyberbullying as these behaviors can have the potential to impact learning (Shariff & Hoff, 2007). When youths are involved in cyberbullying, as perpetrators or victims, they are less likely to perceive their school and teachers positively when compared to uninvolved youths (Bayar & Ucanok, 2012).

Victims of cyberbullying often fear that their classmates might be cyberbullies, which could further disrupt their ability to concentrate on learning in the classroom (Eden, Heiman, & Olenik-Shemesh,

2013). Such disruptions might reduce their academic attainment and performance. Furthermore, lower school commitment and perceptions of a negative school climate increase youths' perpetration of cyberbullying as they feel less connected to their school (Williams & Guerra, 2007).

More training is needed to increase administrators', staffs', and teachers' awareness of cyberbullying. They should work together to develop effective policies at the school level to reduce cyberbullying and implement strategies to handle such incidents. This professional development can increase teachers' self-efficacy for dealing with cyberbullying. When teachers' confidence in their teaching ability increases, they have stronger commitment to their school and desire to learn more about cyberbullying (Eden et al., 2013). Learning more about cyberbullying can increase teachers' awareness of these behaviors and how to deal with cyberbullying, which can reduce youths' involvement in cyberbullying. Teachers who feel more confident in their abilities are more likely to intervene in cyberbullying (Elledge et al., 2013). Intervening in cyberbullying protects adolescents from experiencing these behaviors. Elementary school teachers reported the most motivation for learning about cyberbullying when compared to middle school and high school teachers. This is problematic as cyberbullying involvement increases from elementary school to middle school (Ybarra et al., 2007). Therefore, there is a need for educator training programs aimed at raising awareness of cyberbullying, particularly in the middle school years.

Peers

Peer relationships are important for youths as they help them learn about social norms that govern acceptable and unacceptable behaviors within the peer group. As a consequence of this learning, youths will engage in more of these acceptable behaviors, as dictated by their peers, even if these behaviors are negative. Cyberbullying often involves pre-existing relationships among youths. Therefore, it is not surprising that the best predictor of cyberbullying is the classroom climate in which these behaviors are encouraged (Festl et al., 2013). Believing that one's friends engaged in cyberbullying is a positive predictor of youths' perpetration of cyberbullying (Hinduja & Patchin, 2013). The peer contagion effect might explain these associations. According to this effect, the negative behaviors engaged in by one's friends have the potential to "spread" other youths within their social network (Sijtsema, Ashwin, Simona, & Gina, 2014).

Peer attachment (i.e., belief that peers will be or will not be there for them when they need it) has a relationship to youths' involvement in bullying. Burton et al. (2013) found that lower of peer attachment were associated positively with cyberbullying perpetration and victimization. Other research has revealed that peer rejection increases youths' cyberbullying involvement (Sevcikova et al., 2015; Wright & Li, 2013b). To explain the relationship between peer rejection and cyberbullying involvement, Wright and Li (2012) explained that peer rejection has the potential to trigger negative emotional responses that increase cyberbullying perpetration and victimization. Some research has explored the potential for cyberbullying to promote youths' social standing in their peer group, both online and offline. Given that digital technologies have a prominent role in youths' lives, these technologies can also be used to promote or maintain youths' social standing in the peer group. To investigate this proposal, Wright (2014c) found that perceived popularity (i.e., a reputational type of popularity in the peer group) was related positively to cyberbullying perpetration six months later among adolescents. The literature in this section suggests that it is important to consider the role of peers in youths' cyberbullying involvement.

NEGATIVE CONSEQUENCES OF CYBERBULLYING

The concern with cyberbullying is triggered by its association with various negative social, psychological, behavioral, and academic difficulties that result from youths' exposure to these behaviors. Cyberbullying can disrupt youths' emotional experiences. In the literature, cybervictims report lower levels of global happiness, general school happiness, school satisfaction, family satisfaction, and self-satisfaction (Toledano, Werch, & Wiens, 2015). Cybervictims also indicate that they feel more anger, sadness, and fear when compared to uninvolved youths (Dehue et al., 2008; Machackova, Dedkova, Sevcikova, & Cerna, 2013; Patchin & Hinduja, 2006).

Cyberbullying also negatively effects youths' academic performance. Cyberbullies and cybervictims experience greater academic problems, such as lower motivation for school, poor academic performance, lower academic attainment, and greater school absences (Belae & Hall, 2007; Yousef & Bellamy, 2015). Furthermore, disruptive classroom behaviors, lower grades, and lower test scores are positively associated with youths' cyberbullying perpetration and victimization (Wright, in press).

Cyberbullies and cybervictims report that they experience greater internalizing and externalizing problems (e.g., Mitchell, Ybarra, & Finkelhor, 2007; Patchin & Hinduja, 2006; Wright, 2014b; Ybarra, Diener-West, & Leaf, 2007). When youths are involved in cyberbullying, they report more suicidal thoughts and suicide attempts when compared to uninvolved youths (Bauman, Toomey, & Walker 2013a). Experiencing or perpetrating cyberbullying increases youths' risk of mental health problems (Beckman et al., 2012) and psychiatric and psychosomatic problems (Sourander et al., 2010).

Many studies focused on the relationship between cyberbullying status (victim, bully) and the psychological and behavior consequences do not take into account youths' involvement in traditional forms of bullying. It is important to consider or control for involvement in traditional forms of bullying because these behaviors are highly correlated with cyberbullying (Williams & Guerra, 2007; Wright & Li, 2013b). Taking into account the high correlations between traditional bullying and cyberbullying, Bonanno and colleagues (2013) found that youths involved in cyberbullying experienced greater depressive symptoms and suicidal ideation. In similar researcher, Gradinger and colleagues (2009) and Perren et al. (2012) examined the conjoint effects of both cyberbullying and traditional bullying on youths' psychological and behavioral problems. They found that victims of both traditional face-to-face bullying and cyberbullying reported higher levels of internalizing symptoms when compared to children and adolescents who experienced only one type of victimization. Consequently, these findings indicate that being exposed to different types of bullying behaviors increases youths' experience of depression, anxiety, and loneliness. These findings further support the need for researchers to consider youths' involvement in both offline and online bullying to better understand these relationships and how to best intervene.

SOLUTIONS AND RECOMMENDATIONS

Cyberbullying is a global concern that needs attention by everyone in our communities.

It is important for school curriculum to include elements that teach youths about cyberbullying, digital literacy (i.e., the ability to effectively find, identity, evaluate, and use information obtained from digital content), and digital citizenship (i.e., using information technology in order to engage in society, politics, and government; Cassidy et al., 2012b). This curriculum should also incorporate the positive uses of digital technology, encourage the development of provictim attitudes, empathy, self-esteem, and social skills, and discourage normative beliefs that support bullying behaviors. Schools should also aim

to improve school climate by learning students' names, praising good behavior, and staying technologically up-to-date (Hinduja & Patchin, 2012). It might also be helpful for schools to develop a code of conduct that addressed appropriate digital technology use. Not only is the development of this code of conduct necessary but so is the enforcement of these policies.

Parents have an important role in helping to reduce cyberbullying. They should partner with educators from their children's school and increase their awareness and knowledge of digital technologies and negative online behaviors (Cassidy et al., 2012a; Diamanduros & Downs, 2011). In addition, parents should implement parental monitoring strategies (i.e., strategies parents use to mediate, monitor, and manage their children's relationship with digital media) that help to reduce their children's vulnerability to cyberbullying. They should keep an open dialogue with their children concerning appropriate digital technology use and behaviors. It is also important for parents to be aware of their own digital technology habits, as their children are always watching. Some parents might utilize mobile phones and texting while driving or use mobile phones during dinner or other special events. Consequently, it is important for parents to model appropriate online behavior and technology use to serve as role models for their children.

To fully address cyberbullying, there needs to be multi-pronged approaches and the coordination of multiple stakeholders. It might be difficult for schools and parents to effectively deal with cyberbullying as oftentimes adolescents do not report their experience with these behaviors. In addition, teachers and parents do not always have well-developed technology skills to help when adolescents report experiencing cyberbullying. Parents and teachers might not have the tools for providing effective strategies for dealing with cyberbullying when it occurs and preventing adolescents' exposure.

FUTURE RESEARCH DIRECTIONS

The aim of this chapter was to discuss cyberbullying and the role of different characteristics and risk factors these increase youths' involvement in these behaviors. The literature reviewed in this chapter indicates that there are some noticeable limitations and future directions for research. Anonymity is proposed to have a prominent role in youths' cyberbullying involvement. Despite this proposal, little attention has been given to how anonymous beliefs about the cyber context develop and how these beliefs relate to cyberbullying. It is also important for other research to focus on anonymous and non-anonymous forms of cyberbullying to better understand the motivations underlying these behaviors, and whether victims might experience different adjustment problems based on the form of cyberbullying that they experience. Such a focus is important as non-anonymous cyberbullying might be perpetrated by a known peer and have a stronger influence on youths' depressive symptoms than if the behaviors were perpetrated by someone who is unknown.

Additional research is needed to better understand the long-term impacts of cyberbullying involvement across different age groups. Many studies focus on early and late adolescents, with little attention given to elementary aged children or young adults (Madden et al., 2013; Ybarra et al., 2007). Examining cyberbullying among younger age groups might make it possible to understand the developmental trajectories of online and offline bullying involvement and could also shed light on the temporal ordering of these bullying behaviors. This research might provide valid conclusions about the age that increases youths' vulnerability to cyberbullying. Intervention and prevention programs could be developed which specific consideration to the specific age group identified as at the most risk for cyberbullying involvement.

CONCLUSION

The literature reviewed in this chapter highlights the need for continued research investigations on youths' cyberbullying involvement. The earlier research on cyberbullying aimed to understand the prevalence rates of these behaviors. Although this research provided a solid understanding of the extent of cyberbullying among youths, an understanding of the risk factors and consequences associated with involvement in cyberbullying was a needed focus. Many of the studies on the causes and consequences of cyberbullying utilize concurrent research designs, making it difficult to understand the long-term impact of these behaviors. Furthermore, most of these studies focus on individual level predictors, with more attention needed on contextual predictors, like parents, school, and peers. Ultimately, cyberbullying is a global concern, affecting all aspects of our society. Cyberbullying undermines our ethical and moral values, making it incredibly important that we unite and do our part to diminish youths' involvement in cyberbullying.

REFERENCES

- Ang, R. P. (2016). Cyberbullying: Its prevention and intervention strategies. In D. Sibnath (Ed.), *Child safety, welfare and well-being: Issues and challenges* (pp. 25–38). Springer. doi:10.1007/978-81-322-2425-9_3
- Aoyama, I., Utsumi, S., & Hasegawa, M. (2011). Cyberbullying in Japan: Cases, government reports, adolescent relational aggression and parental monitoring roles. In Q. Li, D. Cross, & P. K. Smith (Eds.), *Bullying in the global playground: Research from an international perspective*. Wiley-Blackwell.
- Aricak, T., Siyahhan, S., Uzunhasanoglu, A., Saribeyoglu, S., Ciplak, S., Yilmaz, N., & Memmedov, C. (2008). Cyberbullying among Turkish adolescents. *Cyberpsychology & Behavior*, 11(3), 253–261. doi:10.1089/cpb.2007.0016 PMID:18537493
- Arslan, S., Savaser, S., Hallett, V., & Balci, S. (2012). Cyberbullying among primary school students in Turkey: Self-reported prevalence and associations with home and school life. *Cyberpsychology, Behavior, and Social Networking*, 15(10), 527–533. doi:10.1089/cyber.2012.0207 PMID:23002988
- Ayas, T., & Horzum, M. B. (2010). *Cyberbullg / victim scale development study*. Retrieved from: <http://www.akademikbakis.org>
- Barlett, C. P., & Gentile, D. A. (2012). Long-term psychological predictors of cyber-bullying in late adolescence. *Psychology of Popular Media Culture*, 2, 123–135. doi:10.1037/a0028113
- Barlett, C. P., Gentile, D. A., Anderson, C. A., Suzuki, K., Sakamoto, A., Yamaoka, A., & Katsura, R. (2013). Cross-cultural differences in cyberbullying behavior: A short-term longitudinal study. *Journal of Cross-Cultural Psychology*, 45(2), 300–313. doi:10.1177/0022022113504622
- Bauman, S., Toomey, R. B., & Walker, J. L. (2013a). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence*, 36(2), 341–350. doi:10.1016/j.adolescence.2012.12.001 PMID:23332116

Bauman, S., Underwood, M. K., & Card, N. A. (2013b). Definitions: Another perspective and a proposal for beginning with cyberaggression. In S. Bauman, D. Cross, & J. Walker (Eds.), *Principles of cyberbullying research: Definitions, measures, methodology* (pp. 26–40). Routledge.

Bayar, Y., & Ucanok, Z. (2012). School social climate and generalized peer perception in traditional and cyberbullying status. *Educational Sciences: Theory and Practice*, 12, 2352–2358.

Beckman, L., Hagquist, C., & Hellstrom, L. (2012). Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional & Behavioural Difficulties*, 17(3-4), 421–434. doi:10.1080/13632752.2012.704228

Bonanno, R. A., & Hymel, S. (2013). Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. *Journal of Youth and Adolescence*, 42(5), 685–697. doi:10.1007/10964-013-9937-1 PMID:23512485

Boulton, M., Lloyd, J., Down, J., & Marx, H. (2012). Predicting undergraduates' self-reported engagement in traditional and cyberbullying from attitudes. *Cyberpsychology, Behavior, and Social Networking*, 15(3), 141–147. doi:10.1089/cyber.2011.0369 PMID:22304402

Brighi, A., Guarini, A., Melotti, G., Galli, S., & Genta, M. L. (2012). Predictors of victimisation across direct bullying, indirect bullying and cyberbullying. *Emotional & Behavioural Difficulties*, 17(3-4), 375–388. doi:10.1080/13632752.2012.704684

Burton, K. A., Florell, D., & Wygant, D. B. (2013). The role of peer attachment and normative beliefs about aggression on traditional bullying and cyberbullying. *Psychology in the Schools*, 50(2), 103–114. doi:10.1002/pits.21663

Cappadocia, M. C., Craig, W. M., & Pepler, D. (2013). Cyberbullying: Prevalence, stability and risk factors during adolescence. *Canadian Journal of School Psychology*, 28(2), 171–192. doi:10.1177/0829573513491212

Cassidy, W., Brown, K., & Jackson, M. (2012a). “Making kind cool”: Parents' suggestions for preventing cyber bullying and fostering cyber kindness. *Journal of Educational Computing Research*, 46(4), 415–436. doi:10.2190/EC.46.4.f

Cassidy, W., Brown, K., & Jackson, M. (2012b). “Under the radar”: Educators and cyberbullying in schools. *School Psychology International*, 33(5), 520–532. doi:10.1177/0143034312445245

Corcoran, L., Connolly, I., & O'Moore, M. (2012). Cyberbullying in Irish schools: An investigation of personality and self-concept. *The Irish Journal of Psychology*, 33(4), 153–165. doi:10.1080/03033910.2012.677995

Curelaru, M., Iacob, I., & Abalasei, B. (2009). *School bullying: Definition, characteristics, and intervention strategies*. Lumea Publishing House.

Dehue, F., Bolman, C., & Vollink, T. (2008). Cyberbullying: Youngsters' experiences and parental perception. *CyberPsychology & Behavior*, 11(2), 217–223. doi:10.1089/cpb.2007.0008 PMID:18422417

Dehue, F., Bolman, C., Vollink, T., & Pouwelse, M. (2012). Cyberbullying and traditional bullying in relation to adolescents' perceptions of parenting. *Journal of Cyber Therapy and Rehabilitation*, 5, 25–34.

- deLara, E. W. (2012). Why adolescents don't disclose incidents of bullying and harassment. *Journal of School Violence, 11*(4), 288–305. doi:10.1080/15388220.2012.705931
- Diamanduros, T., & Downs, E. (2011). Creating a safe school environment: How to prevent cyberbullying at your school. *Library Media Connection, 30*(2), 36–38.
- Eden, S., Heiman, T., & Olenik-Shemesh, D. (2013). Teachers' perceptions, beliefs and concerns about cyberbullying. *British Journal of Educational Technology, 44*(6), 1036–1052. doi:10.1111/j.1467-8535.2012.01363.x
- Elledge, L. C., Williford, A., Boulton, A. J., DePaolis, K. J., Little, T. D., & Salmivalli, C. (2013). Individual and contextual predictors of cyberbullying: The influence of children's provictim attitudes and teachers' ability to intervene. *Journal of Youth and Adolescence, 42*(5), 698–710. doi:10.1007/10964-013-9920-x PMID:23371005
- Fanti, K. A., Demetriou, A. G., & Hawa, V. V. (2012). A longitudinal study of cyberbullying: Examining risk and protective factors. *European Journal of Developmental Psychology, 8*(2), 168–181. doi:10.1080/17405629.2011.643169
- Festl, R., Schwarkow, M., & Quandt, T. (2013). Peer influence, internet use and cyberbullying: A comparison of different context effects among German adolescents. *Journal of Children and Media, 7*(4), 446–462. doi:10.1080/17482798.2013.781514
- Goebert, D., Else, I., Matsu, C., Chung-Do, J., & Chang, J. Y. (2011). The impact of cyberbullying on substance use and mental health in a multiethnic sample. *Maternal and Child Health Journal, 15*(8), 1282–1286. doi:10.1007/10995-010-0672-x PMID:20824318
- Grading, P., Strohmeier, D., & Spiel, C. (2009). Traditional bullying and cyberbullying. *The Journal of Psychology, 217*, 205–213.
- Grigg, D. W. (2012). Definitional constructs of cyberbullying and cyber aggression from a triagnulatory overview: A preliminary study into elements. *Journal of Aggression, Conflict and Peace Research, 4*(4), 202–215. doi:10.1108/17596591211270699
- Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization. *Journal of School Violence, 6*(3), 89–112. doi:10.1300/J202v06n03_06
- Hinduja, S., & Patchin, J. W. (2012). Cyberbullying: Neither and epidemic nor a rarity. *European Journal of Developmental Psychology, 9*(5), 539–543. doi:10.1080/17405629.2012.706448
- Hinduja, S., & Patchin, J. W. (2013). Social influences on cyberbullying behaviors among middle and high school students. *Journal of Youth and Adolescence, 42*(5), 711–722. doi:10.1007/10964-012-9902-4 PMID:23296318
- Huang, Y., & Chou, C. (2010). An analysis of multiple factors of cyberbullying among junior high school students in Taiwan. *Computers in Human Behavior, 26*(6), 1581–1590. doi:10.1016/j.chb.2010.06.005
- Jang, H., Song, J., & Kim, R. (2014). Does the offline bully-victimization influence cyberbullying behavior among youths? Application of general strain theory. *Computers in Human Behavior, 31*, 85–93. doi:10.1016/j.chb.2013.10.007

- Kochenderfer-Ladd, B., & Pelletier, M. (2008). Teachers' views and beliefs about bullying: Influences on classroom management strategies and students' coping with peer victimization. *Journal of School Psychology, 46*(4), 431–453. doi:10.1016/j.jsp.2007.07.005 PMID:19083367
- Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *The Journal of Adolescent Health, 41*(6), 22–30. doi:10.1016/j.jadohealth.2007.08.017 PMID:18047942
- Kwan, G. C. E., & Skoric, M. M. (2013). Facebook bullying: An extension of battles in school. *Computers in Human Behavior, 29*(1), 16–25. doi:10.1016/j.chb.2012.07.014
- Laftman, S. B., Modin, B., & Ostberg, V. (2013). Cyberbullying and subjective health: A large-scale study of students in Stockholm, Sweden. *Children and Youth Services Review, 35*(1), 112–119. doi:10.1016/j.childyouth.2012.10.020
- Lazuras, L., Barkoukis, V., Ourda, D., & Tsorbatzoudis, H. (2013). A process model of cyberbullying in adolescence. *Computers in Human Behavior, 29*(3), 881–887. doi:10.1016/j.chb.2012.12.015
- Lenhart, A. (2015). *Teens, social media & technology overview 2015*. Retrieved from: <https://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/>
- Li, Q. (2007). Bullying in the new playground: Research into cyberbullying and cybervictimization. *Australian Journal of Educational Technology, 23*, 435–454.
- Li, Q. (2008). A cross-cultural comparison of adolescents' experience related to cyberbullying. *Educational Research, 50*(3), 223–234. doi:10.1080/00131880802309333
- Machackova, H., Dedkova, L., Sevcikova, A., & Cerna, A. (2013). Bystanders' support of cyberbullied schoolmates. *Journal of Community & Applied Social Psychology, 23*(1), 25–36. doi:10.1002/casp.2135
- Mason, K. (2008). Cyberbullying: A preliminary assessment for school personnel. *Psychology in the Schools, 45*(4), 323–348. doi:10.1002/pits.20301
- McQuade, C. S., Colt, P. J., & Meyer, B. N. (2009). *Cyber bullying: Protecting kids and adults from online bullies*. Praeger.
- Mitchell, K. J., Ybarra, M., & Finkelhor, D. (2007). The relative importance of online victimization in understanding depression, delinquency, and substance use. *Child Maltreatment, 12*(4), 314–324. doi:10.1177/1077559507305996 PMID:17954938
- Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice, 4*(2), 148–169. doi:10.1177/1541204006286288
- Perren, S., Dooley, J., Shaw, T., & Cross, D. (2010). Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. *Child and Adolescent Psychiatry and Mental Health, 4*(1), 1–10. doi:10.1186/1753-2000-4-28 PMID:21092266
- Pornari, C. D., & Wood, J. (2010). Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior, 36*(2), 81–94. doi:10.1002/ab.20336 PMID:20035548
- Rideout, V. J., Roberts, D. F., & Foehr, U. G. (2005). *Generation M: Media in the lives of 8-18-year-olds: Executive summary*. Henry J. Kaiser Family Foundation.

- Rosen, L. D. (2007). *Me, Myspace, and I: Parenting the Net Generation*. Palgrave Macmillan.
- Sahin, M. (2010). Teachers' perceptions of bullying in high schools: A Turkish study. *Social Behavior and Personality*, 38(1), 127–142. doi:10.2224/bp.2010.38.1.127
- Sevcikova, A., Machackova, H., Wright, M. F., Dedkova, L., & Cerna, A. (2015). Social support seeking in relation to parental attachment and peer relationships among victims of cyberbullying. *Australian Journal of Guidance & Counselling*, 15, 1–13. doi:10.1017/jgc.2015.1
- Shapka, J. D., & Law, D. M. (2013). Does one size fit all? Ethnic differences in parenting behaviors and motivations for adolescent engagement in cyberbullying. *Journal of Youth and Adolescence*, 42(5), 723–738. doi:10.1007/10964-013-9928-2 PMID:23479327
- Shariff, S., & Hoff, D. L. (2007). Cyber bullying: Clarifying legal boundaries for school supervision in cyberspace. *International Journal of Cyber Criminology*, 1, 76–118.
- Sijtsema, J. J., Ashwin, R. J., Simona, C. S., & Gina, G. (2014). Friendship selection and influence in bullying and defending. *Effects of moral disengagement. Developmental Psychology*, 50(8), 2093–2104. doi:10.1037/a0037145 PMID:24911569
- Sjurso, I. R., Fandream, H., & Roland, E. (2016). Emotional problems in traditional and cyber victimization. *Journal of School Violence*, 15(1), 114–131. doi:10.1080/15388220.2014.996718
- Smith, P. K., Del Barrio, C., & Tokunaga, R. S. (2013). Definitions of bullying and cyberbullying: How useful are the terms? In S. Bauman, D. Cross, & J. Walker (Eds.), *Principles of cyberbullying research: Definitions, measures, methodology* (pp. 26–40). Routledge.
- Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 49(4), 376–385. doi:10.1111/j.1469-7610.2007.01846.x PMID:18363945
- Sourander, A., Brunstein, A., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., & Helenius, H. (2010). Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of General Psychiatry*, 67(7), 720–728. doi:10.1001/archgenpsychiatry.2010.79 PMID:20603453
- Stoll, L. C., & Block, R. Jr. (2015). Intersectionality and cyberbullying: A study of cybervictimization in a Midwestern high school. *Computers in Human Behavior*, 52, 387–391. doi:10.1016/j.chb.2015.06.010
- Strohmeier, D., Aoyama, I., Gradinger, P., & Toda, Y. (2013). Cybervictimization and cyberaggression in Eastern and Western countries: Challenges of constructing a cross-cultural appropriate scale. In S. Bauman, D. Cross, & J. L. Walker (Eds.), *Principles of cyberbullying research: Definitions, measures, and methodology* (pp. 202–221). Routledge.
- Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, 7(3), 321–326. doi:10.1089/1094931041291295 PMID:15257832
- Tangen, D., & Campbell, M. (2010). Cyberbullying prevention: One primary school's approach. *Australian Journal of Guidance & Counselling*, 20(2), 225–234. doi:10.1375/ajgc.20.2.225

- Toledano, S., Werch, B. L., & Wiens, B. A. (2015). Domain-specific self-concept in relation to traditional and cyber peer aggression. *Journal of School Violence*, 14(4), 405–423. doi:10.1080/15388220.2014.935386
- Wade, A., & Beran, T. (2011). Cyberbullying: The new era of bullying. *Canadian Journal of School Psychology*, 26(1), 44–61. doi:10.1177/0829573510396318
- Wright, M. F. (2013). The relationship between young adults' beliefs about anonymity and subsequent cyber aggression. *Cyberpsychology, Behavior, and Social Networking*, 16(12), 858–862. doi:10.1089/cyber.2013.0009 PMID:23849002
- Wright, M. F. (2014a). Cyber victimization and perceived stress: Linkages to late adolescents' cyber aggression and psychological functioning. *Youth & Society*.
- Wright, M. F. (2014b). Predictors of anonymous cyber aggression: The role of adolescents' beliefs about anonymity, aggression, and the permanency of digital content. *Cyberpsychology, Behavior, and Social Networking*, 17(7), 431–438. doi:10.1089/cyber.2013.0457 PMID:24724731
- Wright, M. F. (2014c). Longitudinal investigation of the associations between adolescents' popularity and cyber social behaviors. *Journal of School Violence*, 13(3), 291–314. doi:10.1080/15388220.2013.849201
- Wright, M. F. (2015, December). Adolescents' cyber aggression perpetration and cyber victimization: The longitudinal associations with school functioning. *Social Psychology of Education*, 18(4), 653–666. doi:10.1007/11218-015-9318-6
- Wright, M. F. (2015). Cyber victimization and adjustment difficulties: The mediation of Chinese and American adolescents' digital technology usage. *Cyberpsychology (Brno)*, 1(1), 1. <https://cyberpsychology.eu/view.php?cisloclanku=2015051102&article=1>. doi:10.5817/CP2015-1-7
- Wright, M. F., Kamble, S., Lei, K., Li, Z., Aoyama, I., & Shruti, S. (2015). Peer attachment and cyberbullying involvement among Chinese, Indian, and Japanese adolescents. *Societies (Basel, Switzerland)*, 5(2), 339–353. doi:10.3390/soc5020339
- Wright, M. F., & Li, Y. (2012). Kicking the digital dog: A longitudinal investigation of young adults' victimization and cyber-displaced aggression. *Cyberpsychology, Behavior, and Social Networking*, 15(9), 448–454. doi:10.1089/cyber.2012.0061 PMID:22974350
- Wright, M. F., & Li, Y. (2013a). Normative beliefs about aggression and cyber aggression among young adults: A longitudinal investigation. *Aggressive Behavior*, 39(3), 161–170. doi:10.1002/ab.21470 PMID:23440595
- Wright, M. F., & Li, Y. (2013b). The association between cyber victimization and subsequent cyber aggression: The moderating effect of peer rejection. *Journal of Youth and Adolescence*, 42(5), 662–674. doi:10.1007/10964-012-9903-3 PMID:23299177
- Ybarra, M. L., Diener-West, M., & Leaf, P. (2007). Examining the overlap in internet harassment and school bullying: Implications for school intervention. *The Journal of Adolescent Health*, 1(6), 42–50. doi:10.1016/j.jadohealth.2007.09.004 PMID:18047944

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 45(7), 1308–1316. doi:10.1111/j.1469-7610.2004.00328.x PMID:15335350

Yousef, W. S. M., & Bellamy, A. (2015). The impact of cyberbullying on the self-esteem and academic functioning of Arab American middle and high school students. *Electronic Journal of Research in Educational Psychology*, 23(3), 463–482.

Zhou, Z., Tang, H., Tian, Y., Wei, H., Zhang, F., & Morrison, C. M. (2013). Cyberbullying and its risk factors among Chinese high school students. *School Psychology International*, 34(6), 630–647. doi:10.1177/0143034313479692

ADDITIONAL READING

Bauman, S. (2011). *Cyberbullying: What counselors need to know*. American Counseling Association.

Bauman, S., Cross, D., & Walker, J. (2013). *Principles of cyberbullying research: Definitions, measures, and methodology*. Routledge.

Erdur-Baker, O. (2010). Cyberbullying and its correlation to traditional bullying, gender and frequent and risky usage of internet-mediated communication tools. *New Media & Society*, 12(1), 109–125. doi:10.1177/1461444809341260

Hinduja, S., & Patchin, J. W. (2015). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Sage Publications.

Li, Q., Cross, D., & Smith, P. K. (2012). *Cyberbullying in the global playground*. Blackwell Publishing. doi:10.1002/9781119954484

Machackova, H., Dedkova, L., & Mezulanikova, K. (2015). Brief report: The bystander effect in cyberbullying incidents. *Journal of Adolescence*, 43, 96–99. doi:10.1016/j.adolescence.2015.05.010 PMID:26070168

Menesini, E., & Spiel, C. (2012). *Cyberbullying: Development, consequences, risk and protective factors*. Psychology Press.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26(3), 277–287. doi:10.1016/j.chb.2009.11.014

KEY TERMS AND DEFINITIONS

Anonymity: The quality of being unknown or unacknowledged.

Anxiety: A mental health disorder which includes symptoms of worry, anxiety, and/or fear that are intense enough to disrupt one's daily activities.

Collectivism: A cultural value that stressed the importance of the group over individual goals and cohesion within social groups.

Cyberbullying: Children's and adolescents' usage of electronic technologies to hostilely and intentionally harass, embarrass, and intimidate others.

Empathy: The ability to understand or feel what another person is experiencing or feeling.

Externalizing Difficulties: Includes children's and adolescents' failure to control their behaviors.

Individualism: The belief that each person is more important than the needs of the whole group or society.

Loneliness: An unpleasant emotional response to isolation or lack of companionship.

Normative Belief: Beliefs about the acceptability and tolerability of a behavior.

Parental Mediation and Monitoring: The strategies that parents use to manage the relationship between their children and media.

Parenting Style: The standard strategies that parents use in their child rearing.

Peer Attachment: The internalization of the knowledge that their peers will be available and responsive.

Peer Contagion: The transmission or transfer of deviant behavior from one adolescent to another.

Provictim Attitudes: The belief that bullying is unacceptable and that defending victims is valuable.

Social Exclusion: The process involving individuals or groups of people block or deny someone from the group.

Traditional Face-to-Face Bullying: The use of strength or influence to intimidate or physically harm someone.

What Is Cyberbullying?

Michelle F. Wright

Pennsylvania State University, USA & Masaryk University, USA

INTRODUCTION

Children and adolescents spend a great deal of time using and interacting through electronic technologies, including cell phones, gaming consoles, and the Internet (Madden, Lenhart, Duggan, Cortesi, & Gasser, 2013). They engage in and are exposed to different behaviors and content. Some of their involvement with electronic technologies includes many benefits, such as the ability to engage in quick communication with just about anyone, including friends and family, and having access to a multitude of rich information. Despite the many opportunities afforded by electronic technologies, many children and adolescents are exposed to risks, including identify theft, exposure to pedophiles, and viewing unwanted, gory, and/or graphic content. One risk associated with electronic technology usage among children and adolescents is cyberbullying. Cyberbullying occurs through electronic technologies, including gaming consoles, email, instant messaging, chatrooms, social media, and text messages via mobile phones. The attention given to cyberbullying incidences are due to the nature of electronic communications. Through electronic technologies, children and adolescents can remain anonymous, allowing them to harm their victims without experiencing repercussions. The rapid exchange of information and other content through electronic technologies is another concern. For example, a nasty image or video can spread within seconds through electronic technologies. Cyberbullying can also involve multiple bystanders who are also capable of resharing content. The literature in this chapter draws on research from various disciplines, including communication, computer science, education, media studies, psychology, social work, and sociology. Furthermore, the literature involves a variety of different research designs, including cross-sectional and longitudinal methodologies as well as qualitative and quantitative designs. The chapter is organized into the following six sections:

- (1) The first section provides definitions of cyberbullying, the technologies used to target others, the features of anonymity as applied to cyberbullying, and the prevalence rates of children's and adolescents' involvement in cyberbullying.
- (2) The second section describes the individual characteristics and risks associated with children's and adolescents' involvement in cyberbullying.
- (3) The third section details the role of parents and families in children's and adolescents' cyberbullying perpetration and victimization.
- (4) Similar to the third section, the fourth section explains the role of peers and school in children's and adolescents' cyberbullying involvement.
- (5) The purpose of the fifth section is to review literature on the psychological, behavioral, and academic consequences associated with cyberbullying involvement among children and adolescents.
- (6) The final section discusses future research and public policy recommendations related to drawing attention to cyberbullying as an important global health concern.

DOI: 10.4018/978-1-7998-3479-3.ch027

The Nature of Cyberbullying

As deliberately embarrassing or intimidating, cyberbullying involves the usage of modern electronic technologies to harm others using hostile and repetitive behaviors (Kowalski & Limber, 2007; Topcu, Erdur-Backer, & Aydin, 2008; Wolak, Mitchell, & Finkelhor, 2007; Ybarra, West, & Leaf, 2007). Cyberbullying is described as an extension of traditional face-to-face bullying, and it also includes elements of an imbalance of power between the bully and the victim as well as the incorporation of a technological component (Olweus, 1999; Topcu et al., 2008; Ybarra et al., 2007). These behaviors are repetitive, deliberate, and intentionally carried out by bullies with malicious intent (Grigg, 2010). Similar to the definition of traditional face-to-face bullying, cyberbullying can also involve repetition, although this component of the definition is more complex when the behavior involves electronic technologies. For instance, bullies might target victims once by posting humiliating videos but the victimization might continue when bystanders forward the content to others, either with malicious or benign intent. Similar to traditional face-to-face bullying, cyberbullying can also include behaviors with a face-to-face equivalent, such as spreading a rumor about a victim, harassment, physical threats, social exclusion, humiliation, gossiping about a victim to get others not to like the victim, and/or verbal insults. There are also physical forms of cyberbullying, like in traditional face-to-face bullying, which can include hacking. It can include making anonymous phone calls, theft of identity information by pretending to be someone else, distributing explicit videos via various websites, and harassment using instant messenger, social networking websites, and text messages through mobile phones (Wolak et al., 2007; Wright & Li, 2012; Ybarra & Mitchell, 2004). Other forms of cyberbullying involve happy slapping and flaming (Smith et al., 2008). Furthermore, cyberbullying can involve using various electronic technologies, with one study revealing that over half of their sample harmed others through instant messaging tools and social networking websites (Arslan et al., 2012).

Researchers have attempted to understand why children and adolescents engage in cyberbullying. One proposal is that new electronic technologies allow cyberbullies to hide their identities, furthering the power differential between the cyberbully and the cybervictim (Wright, 2013; Ybarra et al., 2007). Being anonymous through electronic technologies involves an imbalance of power between cyberbullies and cybervictims. Many children and adolescents who engage in cyberbullying choose to remain anonymous while perpetrating cyberbullying (Dehue, Bolman, & Vollink, 2008; Kowalski & Limber, 2007; Wright, 2014a; Ybarra & Mitchell, 2004). Another proposal relates to new electronic technologies' ability to allow the cyberbully to perpetrate frequent, repeated, and prolonged anonymous and/or non-anonymous harassment (Wolak et al., 2007; Wright, 2014a; Ybarra & Mitchell, 2004). Cyberbullying might be perpetrated because of the cyberbullies' ability to remain anonymous. Rates of anonymous cyberbullying vary from 12% in one study (Wright, 2014b) to 80% in another study (Arslan et al., 2012). Some research approaches the relationship between anonymity and cyberbullying differently. In this research, Moore et al. (2012) found that more anonymous forums triggered more cyberbullying incidences than less anonymous forums which require login details and passwords. Wright (2013; 2014a) found that when individuals felt more confidence in their ability to remain anonymous while using electronic technologies they engaged in more cyberbullying, especially when they could perpetrate cyberbullying anonymously. Anonymous cyberbullying intensifies cybervictims' negative feelings, including powerlessness (Dooley, Pyzalski, & Cross, 2009; Nocentini, Galmaestra, & Schultze-Krumbholz, 2010).

When an awareness of cyberbullying was gained by researchers, there were many studies conducted to understand how frequently children and adolescents were involved in these behaviors. In 2007, Wolak and colleagues (2007) conducted one of the earliest studies on cyberbullying in the United States. They

What Is Cyberbullying?

found that 50% of children and adolescents in their sample were victimized by cyberbullying. Lower prevalence rates have been found in other studies conducted on cyberbullying involvement in the United States. For instance, Kowalski and Limber (2007) found that 11% of their 3,767 sample from middle school (aged 11-14) were victimized by cyberbullying, 4% had cyberbullied others, and 7% were classified as cyberbullies-cybervictims. Similar estimates were found in a study of cyberbullying by Patchin and Hinduja (2006). Among their sample, they found that 29% reported that they were cybervictims, while 47% reported that they had witnessed cyberbullying at least once. More recent research revealed that 4.9% of children and adolescents in their sample ($N = 4,441$, 6-12th grade) were cybervictims within the past 30 days (Hinduja & Patchin, 2013).

Increasing evidence indicates that cyberbullying is not localized to one country. Instead, research has revealed that cyberbullying is a global concern. In one study, Cappadocia et al. (2013) found that out of their 1,972 sample of Canadian 10th graders, 2.1% perpetrated cyberbullying, 1.9% were cybervictims, and .6% were cyberbullies-cybervictims. Rates of cyberbullying involvement have also been examined in Australia as well. In this research, Campbell and colleagues (2012) found that 4.5% of their sample ($N = 3,112$; grades 6-12th) were victims of cyberbullying. Research has also documented cyberbullying involvement among European children and adolescents. Using a large sample of Swedish adolescents ($N = 22,544$; ages 15-18), Laftman et al. (2013) found that 5% of their sample were classified as cybervictims, 4% as cyberbullies, and 2% as cyberbullies-cybervictims. Rates were similar for cyber victimization in Ireland and Belgium such that 6% of Irish adolescents ($N = 876$; ages 12-17; Corcoran, Connolly, & O'Moore, 2012) and 6.3% of Belgian adolescents ($N = 1,042$, M age = 15.47; Heirman & Walrave, 2012) were classified as cybervictims. Prevalence rates of cyber victimization were much higher among Italian adolescents. For instance, Brighi and colleagues (2012) found that 12.5% of the 2,326 Italian adolescents in their sample were cybervictims. Cyber victimization was reported by 18% of Turkish adolescents in one sample, with 6% classified as cyberbullies ($N = 756$; 7th graders; Yilmaz, 2011). Another study found that 32% of adolescents in one Turkish sample ($N = 276$; ages 14-18) were cybervictims (Erdur-Baker, 2010).

Although slower to develop, some research has focused on cyberbullying involvement rates among children and adolescents in Asia. In one study, Huang and Chou (2010) found that 63.4% of their sample ($N = 545$) witnessed cyberbullying, 34.9% were cybervictims, and 20.4% were cyberbullies. Jang and colleagues (2014) examined rates of cyberbullying involvement among Korean adolescents, with findings revealing that 43% of adolescents in their sample ($N = 3,238$) were cybervictims or perpetrators of cyberbullying. Among a sample of Chinese adolescents ($N = 1,438$), Zhou et al. (2013) found that 34.8% were cyberbullies and 56.9% were cybervictims. Rates were similar among a sample of 1,676 13 through 17 year olds in Singapore such that 59.4% were cybervictims through Facebook and 56.9% were cyberbullies through Facebook (Kwan & Skoric, 2013).

Comparisons of prevalence rates are difficult as the researchers relied on different samples and measurement of cyberbullying. Regardless of the shortcomings associated with studies on prevalence rates of cyberbullying involvement, it is clear that cyberbullying is a concern for children and adolescents around the world. Such a consideration highlights the importance of studying cyberbullying by attempting to understand the risk factors related to children's and adolescents' involvement in cyberbullying as victims, perpetrators, and bystanders. Understanding risks factors associated with cyberbullying involvement can inform prevention efforts to reduce children's and adolescents' involvement in these behaviors.

Characteristics and Risk Factors Associated With Cyberbullying Involvement

3

After much of the research on the prevalence rates of cyberbullying involvement, many researchers directed their attention to understanding the characteristics and risk factors related to children's and adolescents' involvement in cyberbullying. Most of these earlier investigations focused on the role of demographic variables in cyberbullying perpetration and victimization. Age has been examined as a factor related to cyberbullying among children and adolescents. One study revealed that younger adolescents were at a greater risk of cyber victimization, while older adolescents were most frequently classified as cyberbullies (Ayas & Horzum, 2012). Physical forms of cyberbullying (e.g., hacking) were highest in middle school adolescents (Williams & Guerra, 2007). On the other hand, Wade and Beran (2011) found that 9th graders in their sample experienced the greatest risk of cyberbullying involvement when compared to adolescents in middle school. Such conflicting findings from studies focused on age indicated that this variable might be an inconsistent predictor of cyberbullying involvement. Gender is also another inconsistent predictor of cyberbullying involvement among children and adolescents. In particular, some researchers (Li, 2007; Ybarra et al., 2007) have found that boys self-reported more cyberbullying perpetration when compared to girls, with some research revealing that girls experienced more cyber victimization in comparison to boys (Hinduja & Patchin, 2007; Kowalski & Limber, 2007). In contrast, some researchers (e.g., Dehue et al., 2008; Pornari & Wood, 2010) have found that girls were more often the perpetrators of cyberbullying in comparison to boys, while boys were much more likely to be victims of cyberbullying (e.g., Akbulut, Sahih, & Eristi, 2010; Erdur-Baker, 2010; Huang & Chou, 2010). On the other hand, some researchers have found no gender differences in children's and adolescents' cyberbullying involvement (Beran & Li, 2005; Wright & Li, 2013b).

Researchers have focused on the associations between children's and adolescents' involvement in face-to-face bullying and cyberbullying involvement. This literature has revealed that there are positive associations between the perpetration of offline and online bullying, as well as between face-to-face bullying victimization and cyber victimization (Cappadocia et al., 2013; Mitchell et al., 2007; Steffgen, Kong, Pfetsch, & Melzer, 2011; Fanti, Demetriou, & Hawa, 2012). Longitudinal research evidence also suggests that face-to-face victimization is a long-term risk factor for adolescents' involvement in cyberbullying (Wright & Li, 2012).

Some research has been conducted on other characteristics and risk factors associated with cyberbullying involvement among children and adolescents. One risk factor for cyberbullying involvement is internet usage. In this research, cyber victims and cyberbullies typically report higher rates of chatting, emailing, blogging, gaming, and sending instant messages when compared to uninvolved children and adolescents (Smith et al., 2008; Ybarra & Mitchell, 2004). To explain the association between greater internet usage and cyberbullying involvement, Ybarra and colleagues (2007) proposed that cyber victims and cyberbullies are more likely to disclose personal information online, which worsens their exposure to online risks, such as cyberbullying. Thus, frequent internet usage increases children's and adolescents' willingness to disclose private information through electronic technologies. Higher normative beliefs concerning both traditional face-to-face bullying and cyberbullying were also related to cyberbullying perpetration among adolescents (Ang, Tan, & Mansor, 2010; Burton, Florell, & Wygant, 2013; Wright & Li, 2013a). Such findings indicate that cyberbullies view bullying behaviors as normal, which contributes to their perpetration of these behaviors. Defined as children's and adolescents' beliefs that bullying is unacceptable and that defending victims is valued, provictim attitudes are also related to cyberbullying perpetration. In particular, Elledge and colleagues (2013) found that holding lower levels of provictim attitudes were associated with greater cyberbullying perpetration. Other research has revealed that low

self-control, less empathy, and moral disengagement are also contributors to cyberbullying perpetration (Jang et al., 2014; Robson & Witenberg, 2013).

Few studies have focused on the longitudinal risk factors related to cyberbullying involvement. Utilizing a longitudinal design, Fanti and colleagues' (2012) study revealed that media violence exposure, along with callous and unemotional traits, predicted cyberbullying perpetration one year later. Media violence exposure was also a risk factor associated with cyber victimization. Current research has indicated that perceived stress from parents, peers, and academics also increases cyberbullying perpetration one year later among adolescents (Wright, in press). Considering each of the studies examined in this section, it is clear that there are a variety of individual characteristics and risk factors which make children and adolescents vulnerable to cyberbullying involvement. Other research has focused on the role of other individuals in children's and adolescents' involvement in cyberbullying. Parents and families have also been examined as having a role in children's and adolescents' cyberbullying perpetration and victimization.

The Role of Parents and Families

Parental monitoring and parenting styles have been examined in relation to children's and adolescents' involvement in cyberbullying. Mason (2008) found that 30% of adolescents in his sample utilized the Internet for three hours or more daily, with 50% of these adolescents reporting that their parents monitored their activities. Even when parents do monitor their children's online activities, they are usually not effective because some parents lack technological skills and many do not follow-up on the strategies that they have implemented. In addition, many parents in one sample reported that they were not certain of what their children did online and they were less certain on how to discuss the topic of online activities with their children (Rosen, 2007). McQuade and colleagues (2009) found similar results, with 93% of parents indicating that they set rules of the internet, although 37% of their children reported that they were not given specific rules regarding their internet use. Because parents are not able to engage in an active and ongoing dialogue with their children about online activities and considering that experiencing online risks is almost inevitable for children, it is not likely that some children have the skills necessary to effectively deal with problematic situations that arise online. A handful of studies has focused on parents as serving a protective function in mitigating their children's involvement in cyberbullying. Adolescents who believed that their parents would punish them for engaging in cyberbullying engaged in less of these behaviors (Hinduja & Patchin, 2013). Wright (2015b) also found that high parental mediation reduced adolescents' risk of cyberbullying involvement and the negative adjustment difficulties.

Another line of research focuses on the associations between family characteristics and children's and adolescents' cyberbullying involvement. In this research, Ybarra and Mitchell (2004) revealed that lower caregiver monitoring and poor emotional bonds with caregivers were risk factors associated with cyberbullying. They did not find evidence that family income, parental education, and caregiver status (single or married) were related to cyberbullying perpetration or cyber victimization. On the other hand, Aricak and colleagues' (2012) study revealed that parental unemployed increased children's and adolescents' risk of being involved in cyberbullying. Neglectful parenting is also a risk factor associated with cyberbullying perpetration and cyber victimization. In particular, Dehue and colleagues (2012) found that victims and perpetrators of cyberbullying were more likely to report that their parents engaged in neglectful parenting styles when compared to uninvolved children and adolescents. Similarly, their study also indicated that authoritarian parenting style increased children's and adolescents' risk of cyber victimization. More attention should be given to the role of parents and families in children's and adolescents' cyberbullying involvement.

The Role of Schools and Peers

Once children enter formal schooling, they spend most of their time in such settings until they graduate from high school or college. Consequently, some researchers have directed their attention to the linkages of peers and schools to children and adolescents' bullying involvement. Although schools' role in cyberbullying involvement is a bit complicated, some research has focused on how schools, particularly teachers and school climate, impact children's and adolescents' cyberbullying involvement. In one study, many teachers perceived the impact of overt forms of traditional face-to-face bullying as much worse than covert forms of bullying, like relational aggression and some forms of cyberbullying (Sahin, 2010). Cyberbullying and schools' involvement in these incidences is further complicated because many situations involving cyberbullying occur outside of school through electronic technologies. This makes it difficult for teachers to know about such incidences and for them to intervene or provide support to the victim. Additionally, many teachers have not been properly trained on how to deal with and recognize cyberbully behaviors as well as the risk factors associated with these behaviors, furthering hindering schools' ability to deal with these behaviors. Cassidy and colleagues (2012a) found that some of the Canadian teachers in their sample were unfamiliar with some of the newer electronic technologies. Their lack of knowledge made it difficult for them to deal with cyberbullying, even if a victim came to them for help. Many teachers are more willing to promote face-to-face bullying prevention than cyberbullying prevention programs (Tangen & Campbell, 2010). It is important for teachers, school staff, and administrators to understand more about cyberbullying as children's and adolescents' involvement in these behaviors can negatively influence the learning process.

Cybervictims and cyberbullies view their schools and teachers more negatively than uninvolved children and adolescents (Bayar & Ucanok, 2012). A negative school climate and lower school commitment increase cyberbullying perpetration at schools (Ybarra & Mitchell, 2004; Williams & Guerra, 2007). Another problem inhibiting schools' ability to deal with cyberbullying is that there are few implemented policies and programs, even when teachers were concerned with these behaviors (Cassidy, Brown, & Jackson, 2012b). It is important that teachers are concerned about cyberbullying and understand how to prevent it. Teachers' motivations for learning about cyberbullying and their self-efficacy for believing in their ability to deal with these behaviors decreased from elementary school to middle school. Such a finding from Cassidy and colleagues (2012b) is surprising because cyberbullying typically increase in the middle school years (Ybarra et al., 2007). Teachers are more likely to protect students from cyberbullying when they develop awareness of these behaviors (Elledge et al., 2013).

Another feature of the school environment are children's and adolescents' peers who serve as indicators of social norms, including acceptable and unacceptable behaviors. Festl and colleagues (2013) examined the role of peers in adolescents' cyberbullying involvement. Their findings revealed that one of the best predictors of cyberbullying involvement were classrooms in which the highest incidences of cyberbullying occurred. In addition, adolescents who believed that their friends perpetrated cyberbullying were more likely to perpetrate these behaviors as well (Hinduja & Patchin, 2013). Some research has focused on the quality of peer relationships to understand children's and adolescents' involvement in cyberbullying. In particular, Burton and colleagues (2013) found that higher levels of peer attachment were related negatively to cyberbullying perpetration and cyber victimization. In addition, being rejected by one's peers was also related to cyberbullying, even after controlling for face-to-face victimization (Wright & Li, 2013b). Research has also revealed that cyberbullying perpetration might be used to boost adolescents' social status among their peer group. For example, Wright's (2014b) study suggested that higher levels of perceived popularity, a reputational type of popularity usually linked to traditional

face-to-face bullying perpetration, increased adolescents' cyberbullying perpetration six months later. In another study, Wright (2015b) found that friends served a protective function in reducing adolescents' cyber victimization and the negative adjustment difficulties associated with this experience. Friends might discuss technology skills, online safety, and reinforce appropriate online practices (e.g., not sharing passwords, discussing personal information with strangers) and behaviors.

Psychological, Behavioral, and Academic Difficulties

Researchers, educators, and parents are particularly concerned about the psychological, behavioral, and academic difficulties associated with adolescents experiencing and/or perpetrating cyberbullying. Cybervictims reported lower global happiness, school happiness, school satisfaction, school satisfaction, family satisfaction, and self-satisfaction when compared to uninvolved adolescents (Navarro et al., 2013). Similarly, researchers (e.g., Beran & Li, 2005; Dehue et al., 2008; Patchin & Hinduja, 2006) have found that cybervictims felt greater levels of anger, fearfulness, and sadness in comparison to uninvolved children and adolescents, as well as more emotional and peer problems. Perpetrators and victims of cyberbullying reported internalizing difficulties, such as depression and anxiety, and externalizing difficulties, including drug and alcohol use as well as violence (Arslan et al., 2012; Li, 2007; Patchin & Hinduja, 2012; Ybarra et al., 2007). Cyberbullying involvement is also linked to problems in school. In one study, Wright (2015a) examined the longitudinal associations between school functioning and cyberbullying involvement. She found that cyberbullying perpetration and cyber victimization were each associated with poor academic performance, absenteeism, and school behavioral problems (e.g., classroom disruptions) one year later. Some researchers have focused on psychological and behavioral difficulties associated with the joint effects of children's and adolescents' involvement in traditional face-to-face bullying and cyberbullying. Victims who were victimized by both traditional face-to-face bullying and cyberbullying reported more internalizing difficulties when compared to adolescents who experienced only one type of victimization (Grading, Strohmeier, & Spiel, 2009; Perren, Dooley, Shaw, & Cross, 2010).

The bulk of the research on the psychological, academic, and behavioral difficulties associated with children's and adolescents' cyberbullying involvement do not control for traditional face-to-face bullying involvement. Controlling for these variables is important as there are strong correlations between cyberbullying involvement and traditional face-to-face bullying involvement. To this end, Bonanno and colleagues (2013) examined cyberbullying perpetration and cyber victimization in relation to depressive symptoms and suicidal ideation. After including face-to-face bullying and victimization in the same model, they still found associations between these variables. However, Brighi and colleagues (2012) did not find a relationship between cyber victimization and loneliness or self-esteem, after accounting for traditional face-to-face bullying victimization.

RECOMMENDATIONS

There are a variety of gaps in the cyberbullying literature, which need to be investigated through future research. One recommendation is that follow-up research needs to focus on understanding more about the concept of anonymity as applied to cyberbullying. Anonymity is not only a component of cyberbullying, but anonymous forms of traditional face-to-face bullying can occur as well. For example, a nasty note can be slipped into a peer's locker anonymously. Thus, research should be conducted to understand

differences in children's and adolescents' conceptualizations of anonymous acts in the cyber context and the offline world. There are few longitudinal studies conducted on cyberbullying involvement. In future research, longitudinal designs should be utilized to examine both traditional face-to-face bullying and cyberbullying involvement from childhood into late adolescence or even emerging adulthood. Investigations focused on the developmental trajectory of bullying involvement in both social contexts are important as much of the research on cyberbullying does not take into account previous levels of traditional face-to-face bullying, although these variables are strongly correlated (Wright & Li, 2013a; Ybarra et al., 2007). Researchers are beginning to understand that attacks which occur in cyberspace might be just as powerful as face-to-face attacks. Our communities need to recognize this threat and devote time to dealing with it, as it affects everyone. Cyberbullying will probably not disappear anytime soon, due to our increasing reliance of electronic technologies. Therefore, we all must stand united in our efforts to reduce or prevent cyberbullying.

REFERENCES

- Akbulut, Y., Sahin, T. L., & Eristi, B. (2010). Cyberbullying victimization among Turkish online social utility members. *Journal of Educational Technology & Society*, 13, 192–201.
- Ang, R. P., Kit-Aun, T., & Mansor, A. T. (2010). Normative beliefs about aggression as a mediator of narcissistic exploitativeness and cyberbullying. *Journal of Interpersonal Violence*, 26(13), 2619–2634. doi:10.1177/0886260510388286 PMID:21156699
- Aricak, T., Siyahhan, S., Uzunhasanoglu, A., Saribeyoglu, S., Ciplak, S., Yilmaz, N., & Memmedov, C. (2008). Cyberbullying among Turkish adolescents. *Cyberpsychology & Behavior*, 11(3), 253–261. doi:10.1089/cpb.2007.0016 PMID:18537493
- Arslan, S., Savaser, S., Hallett, V., & Balci, S. (2012). Cyberbullying among primary school students in Turkey: Self-reported prevalence and associations with home and school life. *Cyberpsychology, Behavior, and Social Networking*, 15(10), 527–533. doi:10.1089/cyber.2012.0207 PMID:23002988
- Ayas, T., & Horzum, M. B. (2010). *Cyberbullg / victim scale development study*. Retrieved from: <http://www.akademikbakis.org>
- Bayar, Y., & Ucanok, Z. (2012). School social climate and generalized peer perception in traditional and cyberbullying status. *Educational Sciences: Theory and Practice*, 12, 2352–2358.
- Beran, T., & Li, Q. (2005). Cyber-harassment: A new method for an old behavior. *Journal of Educational Computing Research*, 32(3), 265–277. doi:10.2190/8YQM-B04H-PG4D-BLLH
- Bonanno, R. A., & Hymel, S. (2013). Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. *Journal of Youth and Adolescence*, 42(5), 685–697. doi:10.1007/10964-013-9937-1 PMID:23512485
- Brighi, A., Guarini, A., Melotti, G., Galli, S., & Genta, M. L. (2012). Predictors of victimisation across direct bullying, indirect bullying and cyberbullying. *Emotional & Behavioural Difficulties*, 17(3-4), 375–388. doi:10.1080/13632752.2012.704684

What Is Cyberbullying?

- Burton, K. A., Florell, D., & Wygant, D. B. (2013). The role of peer attachment and normative beliefs about aggression on traditional bullying and cyberbullying. *Psychology in the Schools, 50*(2), 103–114. doi:10.1002/pits.21663
- Campbell, M., Spears, B., Slee, P. H., Butler, D., & Kift, S. (2012). Victims' perceptions of traditional and cyberbullying, and the psychosocial correlates of their victimisation. *Emotional & Behavioural Difficulties, 17*(3-4), 389–401. doi:10.1080/13632752.2012.704316
- Cappadocia, M. C., Craig, W. M., & Pepler, D. (2013). Cyberbullying: Prevalence, stability and risk factors during adolescence. *Canadian Journal of School Psychology, 28*(2), 171–192. doi:10.1177/0829573513491212
- Cassidy, W., Brown, K., & Jackson, M. (2012a). "Making kind cool": Parents' suggestions for preventing cyber bullying and fostering cyber kindness. *Journal of Educational Computing Research, 46*(4), 415–436. doi:10.2190/EC.46.4.f
- Cassidy, W., Brown, K., & Jackson, M. (2012b). "Under the radar": Educators and cyberbullying in schools. *School Psychology International, 33*(5), 520–532. doi:10.1177/0143034312445245
- Corcoran, L., Connolly, I., & O'Moore, M. (2012). Cyberbullying in Irish schools: An investigation of personality and self-concept. *The Irish Journal of Psychology, 33*(4), 153–165. doi:10.1080/03033910.2012.677995
- Dehue, F., Bolman, C., & Vollink, T. (2008). Cyberbullying: Youngsters' experiences and parental perception. *CyberPsychology & Behavior, 11*(2), 217–223. doi:10.1089/cpb.2007.0008 PMID:18422417
- Dehue, F., Bolman, C., Vollink, T., & Pouwelse, M. (2012). Cyberbullying and traditional bullying in relation to adolescents' perceptions of parenting. *Journal of Cyber Therapy and Rehabilitation, 5*, 25–34.
- Dooley, J. J., Pyzalski, J., & Cross, D. (2009). Cyberbullying versus face-to-face bullying. *The Journal of Psychology, 217*, 182–188.
- Elledge, L. C., Williford, A., Boulton, A. J., DePaolis, K. J., Little, T. D., & Salmivalli, C. (2013). Individual and contextual predictors of cyberbullying: The influence of children's provictim attitudes and teachers' ability to intervene. *Journal of Youth and Adolescence, 42*(5), 698–710. doi:10.1007/10964-013-9920-x PMID:23371005
- Erdur-Baker, O. (2010). Cyberbullying and its correlation to traditional bullying, gender and frequent and risky usage of internet-mediated communication tools. *New Media & Society, 12*(1), 109–125. doi:10.1177/1461444809341260
- Fanti, K. A., Demetriou, A. G., & Hawa, V. V. (2012). A longitudinal study of cyberbullying: Examining risk and protective factors. *European Journal of Developmental Psychology, 8*(2), 168–181. doi:10.1080/17405629.2011.643169
- Festl, R., Schwarkow, M., & Quandt, T. (2013). Peer influence, internet use and cyberbullying: A comparison of different context effects among German adolescents. *Journal of Children and Media, 7*(4), 446–462. doi:10.1080/17482798.2013.781514
- Grading, P., Strohmeier, D., & Spiel, C. (2009). Traditional bullying and cyberbullying. *The Journal of Psychology, 217*, 205–213.

- Grigg, D. W. (2010). Cyber-aggression: Definition and concept of cyberbullying. *Australian Journal of Guidance & Counselling*, 20(2), 143–156. doi:10.1375/ajgc.20.2.143
- Heirman, W., & Walrave, M. (2012). Predicting adolescent perpetration in cyberbullying: An application of the theory of planned behavior. *Psicothema*, 24, 614–620. PMID:23079360
- Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3), 206–221. doi:10.1080/13811118.2010.494133 PMID:20658375
- Hinduja, S., & Patchin, J. W. (2012). Cyberbullying: Neither and epidemic nor a rarity. *European Journal of Developmental Psychology*, 9(5), 539–543. doi:10.1080/17405629.2012.706448
- Hinduja, S., & Patchin, J. W. (2013). Social influences on cyberbullying behaviors among middle and high school students. *Journal of Youth and Adolescence*, 42(5), 711–722. doi:10.1007/10964-012-9902-4 PMID:23296318
- Huang, Y., & Chou, C. (2010). An analysis of multiple factors of cyberbullying among junior high school students in Taiwan. *Computers in Human Behavior*, 26(6), 1581–1590. doi:10.1016/j.chb.2010.06.005
- Jang, H., Song, J., & Kim, R. (2014). Does the offline bully-victimization influence cyberbullying behavior among youths? Application of general strain theory. *Computers in Human Behavior*, 31, 85–93. doi:10.1016/j.chb.2013.10.007
- Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *The Journal of Adolescent Health*, 41(6), 22–30. doi:10.1016/j.jadohealth.2007.08.017 PMID:18047942
- Laftman, S. B., Modin, B., & Ostberg, V. (2013). Cyberbullying and subjective health: A large-scale study of students in Stockholm, Sweden. *Children and Youth Services Review*, 35(1), 112–119. doi:10.1016/j.childyouth.2012.10.020
- Li, Q. (2007). Bullying in the new playground: Research into cyberbullying and cybervictimization. *Australian Journal of Educational Technology*, 23, 435–454.
- Madden, M., Lenhart, A., Duggan, M., Cortesi, S., & Gasser, U. (2013). *Teens and technology 2013*. Retrieved from: <https://www.pewinternet.org/2013/03/13/teens-and-technology-2013/>
- Mason, K. (2008). Cyberbullying: A preliminary assessment for school personnel. *Psychology in the Schools*, 45(4), 323–348. doi:10.1002/pits.20301
- McQuade, C. S., Colt, P. J., & Meyer, B. N. (2009). *Cyber bullying: Protecting kids and adults from online bullies*. Praeger.
- Mitchell, K. J., Ybarra, M., & Finkelhor, D. (2007). The relative importance of online victimization in understanding depression, delinquency, and substance use. *Child Maltreatment*, 12(4), 314–324. doi:10.1177/1077559507305996 PMID:17954938
- Moore, M. J., Nakano, T. N., Enomoto, A., & Suda, T. (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, 28(3), 861–867. doi:10.1016/j.chb.2011.12.005

What Is Cyberbullying?

- Navarro, R., Serna, C., Martinez, V., & Ruiz-Oliva, R. (2013). The role of Internet use and parental mediation on cyberbullying victimization among Spanish children from rural public schools. *European Journal of Psychology of Education*, 28(3), 725–745. doi:10.1007/10212-012-0137-2
- Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. *Australian Journal of Guidance & Counselling*, 20(2), 129–142. doi:10.1375/ajgc.20.2.129
- Olweus, D. (1999). Sweden. In K. Smith, Y. Morita, J. Junger-Tas, D. Olweus, R. Catalano, & P. Slee (Eds.), *The nature of school bullying: A cross-national perspective* (pp. 7–27). Routledge.
- Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2), 148–169. doi:10.1177/1541204006286288
- Perren, S., Dooley, J., Shaw, T., & Cross, D. (2010). Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. *Child and Adolescent Psychiatry and Mental Health*, 4(1), 1–10. doi:10.1186/1753-2000-4-28 PMID:21092266
- Pornari, C. D., & Wood, J. (2010). Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior*, 36(2), 81–94. doi:10.1002/ab.20336 PMID:20035548
- Robson, C., & Witenberg, R. T. (2013). The influence of moral disengagement, morally based self-esteem, age, and gender on traditional bullying and cyberbullying. *Journal of School Violence*, 12(2), 211–231. doi:10.1080/15388220.2012.762921
- Rosen, L. D. (2007). *Me, Myspace, and I: Parenting the Net Generation*. Palgrave Macmillan.
- Sahin, M. (2010). Teachers' perceptions of bullying in high schools: A Turkish study. *Social Behavior and Personality*, 38(1), 127–142. doi:10.2224/bp.2010.38.1.127
- Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 49(4), 376–385. doi:10.1111/j.1469-7610.2007.01846.x PMID:18363945
- Steffgen, G., König, A., Pfetsch, J., & Melzer, A. (2011). Are cyberbullies less empathic? Adolescents' cyberbullying behavior and empathic responsiveness. *Cyberpsychology, Behavior, and Social Networking*, 14(11), 643–648. doi:10.1089/cyber.2010.0445 PMID:21554126
- Tangen, D., & Campbell, M. (2010). Cyberbullying prevention: One primary school's approach. *Australian Journal of Guidance & Counselling*, 20(2), 225–234. doi:10.1375/ajgc.20.2.225
- Topcu, C., Erdur-Baker, O., & Capa, A. Y. (2008). Examination of cyber-bullying experiences among Turkish students from different school types. *Cyberpsychology & Behavior*, 11(6), 644–648. doi:10.1089/cpb.2007.0161 PMID:18783345
- Vandebosch, H., & van Cleemput, K. (2008). Defining cyberbullying: A qualitative research into the perceptions of youngsters. *Cyberpsychology & Behavior*, 11(4), 499–503. doi:10.1089/cpb.2007.0042 PMID:18721100
- Wade, A., & Beran, T. (2011). Cyberbullying: The new era of bullying. *Canadian Journal of School Psychology*, 26(1), 44–61. doi:10.1177/0829573510396318

- Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of internet bullying. *The Journal of Adolescent Health*, 41(6), S14–S21. doi:10.1016/j.jadohealth.2007.08.018 PMID:18047941
- Wolak, J., Mitchell, K., & Finkelhor, D. (2007). Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *The Journal of Adolescent Health*, 41(6), 51–58. doi:10.1016/j.jadohealth.2007.08.019 PMID:18047945
- Wright, M. F. (2013). The relationship between young adults' beliefs about anonymity and subsequent cyber aggression. *Cyberpsychology, Behavior, and Social Networking*, 16(12), 858–862. doi:10.1089/cyber.2013.0009 PMID:23849002
- Wright, M. F. (2014a). Predictors of anonymous cyber aggression: The role of adolescents' beliefs about anonymity, aggression, and the permanency of digital content. *Cyberpsychology, Behavior, and Social Networking*, 17(7), 431–438. doi:10.1089/cyber.2013.0457 PMID:24724731
- Wright, M. F. (2014b). Longitudinal investigation of the associations between adolescents' popularity and cyber social behaviors. *Journal of School Violence*, 13(3), 291–314. doi:10.1080/15388220.2013.849201
- Wright, M. F. (2015, November). Cyber victimization and perceived stress: Linkages to late adolescents' cyber aggression and psychological functioning. *Youth & Society*, 47(6), 789–810. doi:10.1177/0044118X14537088
- Wright, M. F. (2015a). Adolescents' cyber aggression perpetration and cyber victimization: The longitudinal associations with school functioning. *Social Psychology of Education*, 18(4), 653–666. doi:10.1007/11218-015-9318-6
- Wright, M. F. (2015b). Cyber victimization and adjustment difficulties: The mediation of Chinese and American adolescents' digital technology usage. *Cyberpsychology (Brno)*, 9(1), 7. doi:10.5817/CP2015-1-7
- Wright, M. F., & Li, Y. (2012). Kicking the digital dog: A longitudinal investigation of young adults' victimization and cyber-displaced aggression. *Cyberpsychology, Behavior, and Social Networking*, 15(9), 448–454. doi:10.1089/cyber.2012.0061 PMID:22974350
- Wright, M. F., & Li, Y. (2013a). Normative beliefs about aggression and cyber aggression among young adults: A longitudinal investigation. *Aggressive Behavior*, 39(3), 161–170. doi:10.1002/ab.21470 PMID:23440595
- Wright, M. F., & Li, Y. (2013b). The association between cyber victimization and subsequent cyber aggression: The moderating effect of peer rejection. *Journal of Youth and Adolescence*, 42(5), 662–674. doi:10.1007/10964-012-9903-3 PMID:23299177
- Ybarra, M. L., Diener-West, M., & Leaf, P. (2007). Examining the overlap in internet harassment and school bullying: Implications for school intervention. *The Journal of Adolescent Health*, 1(6), 42–50. doi:10.1016/j.jadohealth.2007.09.004 PMID:18047944
- Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 45(7), 1308–1316. doi:10.1111/j.1469-7610.2004.00328.x PMID:15335350

What Is Cyberbullying?

Zhou, Z., Tang, H., Tian, Y., Wei, H., Zhang, F., & Morrison, C. M. (2013). Cyberbullying and its risk factors among Chinese high school students. *School Psychology International*, 34(6), 630–647. doi:10.1177/0143034313479692

How Theoretical Frameworks Inform the Understanding of the Relationship Between Gender and Cyberbullying

Monica Bixby Radu

Southeast Missouri State University, USA

Alexandria L. Rook

Southeast Missouri State University, USA

INTRODUCTION

Over the past two decades, bullying has received considerable attention, with scholars, educators, parents, and youths expressing their concerns regarding bullying at school. While studies suggest that bullying is relatively common in elementary, middle, and high schools, bullying also occurs outside of schools. The Internet, as well as other forms of technology, provide a platform that allows bullying to extend beyond face-to-face peer interaction. Scholars identify this form of bullying as *cyberbullying*. Cyberbullying refers to intentional, aggressive behaviors that are perpetrated through mediated communication, such as email, text messages, social media, or other online websites (Sticca & Perren, 2013). Cyberbullying not only affects school-age youth and adolescents, as recent literature suggests that technological advances have also contributed to an increase in workplace cyberbullying among adults (D'Souza, Forsyth, Tappin & Catley, 2018; Vranjes, 2018).

While both bullying and cyberbullying have gained national attention, research remains uncertain regarding the link between gender and various forms of bullying. Seals and Young (2003) suggest that like most crime, males are more likely than females to both perpetrate and be the victim of bullying. Yet, other researchers argue that both males and females are equally likely to engage in bullying, but there are gender differences in the ways in which boys and girls engage in bullying (Day & Kahle, 2014). Other research specifically focuses on gender differences in regard to cyberbullying perpetration and victimization. For example, Li (2006) found that among a sample of 256 youths, 22% of males reported engaging in cyberbullying as perpetrators, compared to 11.6% of female youths. However, there was little difference between male (25%) and female (25.6%) youths in relation to being the *victim* of cyberbullying (Li, 2006). More recently, using an online questionnaire to ask youths about their experiences with cyberbullying (N=208), Wong, Cheung, and Xia (2018) found that compared to female youths, males were more likely to be both the perpetrators *and* victims of cyberbullying. While these studies are valuable contributions to research on cyberbullying, they also suggest that research is inconclusive concerning the relationship between gender and cyberbullying perpetration and victimization.

Scholars recognize the importance of understanding the relationship between gender and various forms of bullying because there may be unique consequences for male and female victims of cyberbullying. For example, drawing from a sample of 1,222 youths across 15 schools in the United States, Carbone-Lopez, Esbensen, and Brick (2010) found that girls tended to experience a broader range of consequences related bullying than their male counterparts, including more severe psychological and health problems.

DOI: 10.4018/978-1-7998-3479-3.ch028

The authors also found that being the victim of indirect or relational bullying had a negative effect on girls' self-esteem, while for boys, this form of bullying was not associated with variations in self-esteem.

Therefore, in this article, the authors argue that to better (1) help individuals cope with their experiences with cyberbullying victimization and (2) prevent individuals from engaging in cyberbullying, scholars, educators, and policymakers need to better understand patterns and trends associated with online bullying. This article contributes to research on gender and cyberbullying by merging insights from criminology, sociology, and gender studies to review prior literature on bullying and more specifically, cyberbullying. Additionally, the authors present how theoretical perspectives, including the "doing gender" perspective and an intersectional approach may help shape our understanding regarding the relationship between gender and cyberbullying. Following this, the authors consider implications for policy development and future research.

BACKGROUND

Scholars identify bullying as intentional, repeated actions meant to harm a victim and perpetuate an inequity of power (Olweus, 1991). Traditional forms of bullying take place in-person, typically occurring in a specific location, such as schools and workplaces. Bullying encompasses both overt and relational forms of abuse and harassment manifested in a multitude of ways, including physical altercations, taunts, gossip, or exclusion from specific groups or certain activities that cause relational damage (Thomas, Connor, & Scott, 2014). Research on traditional bullying distinguish that these actions occur without the assistance of technology; the harassment is often face-to-face or, in the case of gossip, circulated directly from person to person.

Over the last decade, bullying has evolved alongside the technological revolution of computers, cell phones and other communication devices. Deliberate, overt or covert behaviors perpetrated through digital technology is referred to as *cyberbullying* (Mendez, Bauman, & Guillory, 2012). Like traditional forms of bullying, cyberbullying also takes several forms including sending hurtful or threatening text messages or emails, spreading rumors through emails and private messaging, forwarding private pictures to unintended recipients, and posting derogatory comments and pictures on social media. Cyberbullying may take place anonymously or the bully and victim may know each other as peers or even co-workers (Vranjes et al., 2018).

The incidence of cyberbullying has increased as the Internet and communication technologies have grown (Aivazpour & Beebe, 2018). Following this, there has been a surge of research on cyberbullying (Zych, Farrington, & Llorent, 2018). One example is Carter and Wilson (2015) who surveyed a sample of 367 adolescents from the Midwest, asking about their participation in cyberbullying. The authors found that among their sample, 17% reported that they had been the victim of cyberbullying. Carter and Wilson concluded that the already high prevalence of cyberbullying would continue to rise with technological advances. Another example is Wilton and Campbell (2011) who explored the reason why youths engage in cyberbullying. They found that adolescents were more likely to partake in the virtual harassment of their peers for reasons of emotional self-satisfaction. The authors argued that virtual or online bullying may become a more popular alternative to traditional bullying as it allows perpetrators to disassociate from "real" life and provides bullies with a sense of anonymity.

Recently, Gofin and Avitzour (2012) argue the Internet has the ability to connect both friends and complete strangers, and it may blur the lines between actions and direct negative outcomes. Consequently, cyberbullies may not perceive their online behaviors as bullying because of the mediated communication

between online bullying and victim (Gofin & Avitzour, 2012). Therefore, research may be underestimating the prevalence of cyberbullying if the cyberbullies themselves do not realize they are engaging in this problematic behavior.

Consequences Associated with Bullying and Cyberbullying Victimization

Studies consistently find that there are numerous negative consequences associated with being the victim of bullying. For example, being the victim of bullying is associated with lower grades (Konishi, Hymel, Zumbo, & Li, 2010) and avoidance behaviors at school (Hutzell & Payne, 2012). Other studies suggest that victims of bullying are more likely to experience mental health issues than their peers who were not bullied, including higher rates of depression, loneliness, and anxiety (Nishina, Juvonen, & Witkow, 2005). Victims of bullying often have low levels of self-efficacy and more psychosocial adjustment problems than individuals who were not bullied (Nansel et al., 2004). Drawing from longitudinal data, Radu (2018a) found that victims of childhood bullying were more likely to engage in subsequent violent behavior than their non-bullied peers. Bullying may also have a negative effect on educational outcomes, as victims of bullying tend to have higher rates of dropping out of high school compared to students who had not experienced being bullied (Cooper & Nickerson, 2013).

There are numerous adverse mental health consequences associated with cyberbullying victimization (Abreu & Kenny, 2018), including increased feelings of anxiety, depression, and feelings of loneliness, and lower levels of self-esteem (Motswi & Mashegoane, 2017). Victims of cyberbullying often report emotional issues, such as anger, sadness, and depression (Balakrishnam, 2018). In a study comparing cyberbully victims with non-victims, Gorzig (2016) found that victims of cyberbullying were more likely than non-victims to view suicide and self-harm related materials on the Internet. Additional studies suggest that workplace cyberbullying has severe consequences for victims, as Kowalski, Toth, and Morgan (2018) found that victims of cyberbullying reported lower levels of self-esteem and higher levels of social anxiety and depression compared to those who had been the victim of workplace cyberbullying. Motswi & Mashegoane (2017) argue that the mobility of cyberbullying means that individuals can be bullied in several different contexts, which could include home, school, and the workplace, which may increase the risk of long-term negative consequences. Additionally, the negative impact of cyberbullying victimization may be exacerbated by the large audience that the web provides (Slonje & Smith, 2008), meaning there are often bystanders or witnesses of cyberbullying incidents. Online bystanders have the potential to defend or support the bully victim, but research suggest that bystanders may also reinforce or join in cyberbullying the target (Shultz et al., 2014).

Sociological and criminological theories are useful in explaining problematic behaviors, such as experiences with victimization. In the following section, the authors review the value of considering West and Zimmerman's (1987) "doing gender" framework and Crenshaw's (1989) intersectionality to better understand the relationship between gender and cyberbullying.

MAIN FOCUS OF THE CHAPTER

Doing Gender, Bullying, and Cyberbullying

In West and Zimmerman's (1987) seminal work, they argued that gender is a social construct, a routine accomplishment that is embedded in everyday interactions. They argued that gender is produced and

reproduced by social interaction and people are judged in terms of how consistent they are with social beliefs about gender. It is further argued that gender is a situated accomplishment in terms of “normative” expectations for a particular sex category (West & Zimmerman, 1995). West and Zimmerman’s (1987) “doing gender” developed from Goffman’s (1976) concept of gender display, which maintains that displays of gender are highly conventionalized behavioral-exchanges. Within these social exchanges, the presence or absence of symmetry can establish either deference or dominance. Goffman argued that gender is a socially scripted dramatization of the culture’s idealization of feminine and masculine nature, suggesting that gender permeates all social situations; therefore, avoidance of being judged as either masculine or feminine is unavoidable. Gender display focuses on the behavioral aspects of being a woman or man, opposed to biological differences (Goffman, 1976). Drawing from Goffman’s ideas, West and Zimmerman also argued that “doing gender” is unavoidable because gender is an ideological device which produces, reproduces, and legitimates individuals’ choices, as people are often judged regarding how consistent their behaviors are with social standards of their gender. The social arrangements based on one’s sex category are viewed as legitimate ways of organizing social life—men doing dominance and women doing deference—which creates and maintains hierarchical arrangements through what are considered natural or biological differences (West & Zimmerman, 1987).

Understanding how youth “do gender” is important to the study of bullying and cyberbullying because studies consistently find that gender is linked to bullying in several ways. Crick and Grotpeter (1995) found that girls were more likely than boys to engage in covert or relational bullying, findings that are supported by a range of other studies (e.g., Cullerton-Sen & Crick, 2005; Wang, Iannotti, & Nansel, 2009). Crick and Grotpeter (1995) argued that girls’ use of relational bullying tended to be focused on damaging reputations rather than inflicting physical harm, while male students were more likely to be involved in direct, overt, or physical bullying. SooHoo (2009) found that girls were more likely to utilize relational bullying as they attempted to maintain their social status or lower a peer’s status. While Connell (2005) argued that male-perpetrated violence towards other males may be used to display one’s masculinity. Boys and young men may attempt to gain prestige and a positive “tough guy” image by engaging in physical violence towards their male peers. Using in-depth interviews and comparison groups, Messerschmidt’s (2000) reported similar findings. He discovered that masculine identities were constructed through interaction with the home and school, and different “masculine resources” are available for different individuals depending on their circumstances. Messerschmidt found that schools were environments where young men constructed violent masculinities by victimizing their weaker male peers.

Other studies suggest that differences in type of technology use by gender may contribute to both the acceptance and perpetration of cyberbullying. Drawing from a three-wave longitudinal study of adolescents (N = 3,000), Barlett and colleagues (2018) found that female adolescents tended to spend more time on social media websites both on their phones and on their computers than their male peers, and they were more likely to use blogs to express their feelings than male adolescents. However, the authors found that even though female youth spent more time on social media websites than male youth, the male participants were more likely to have positive cyberbullying attitudes and engage in cyberbullying perpetration than their female counterparts. This suggests that while there may be gender differences in technology medium and use, the relationship between gender, technology, and cyberbullying perpetration is complex and additional research is needed to provide a more nuanced understanding of this relationship.

Additionally, there is disagreement within academic literature whether gender is a significant factor in cyberbullying perpetration and victimization. Motswi and Mashegoane (2017) found in a study of 324 secondary school students in the Limpopo Province, Africa that there were more female students affected by cyberbullying (11%) than males (4%). Other studies suggest that suggests that girls tend to

engage in cyberbully to impress boys or to gain male attention by taking out “competitors,” and girls are more likely to insult appearances than resort to physical violence (Goddard, 2008). Drawing from a two-wave panel survey among 1,817 high school students in southwest Germany, Festl and Quandt (2016) found that compared to their female counterparts, male participants were more likely to participate in various forms of cyberbullying as the cyberbullies. Additionally, boys reported overall higher frequencies of cyberbullying victimization experiences than their female peers, including being insulted on a public website, having an embarrassing picture or video of themselves posted online, and receiving a message from someone using a fake identity. These discrepancies regarding the association between gender and cyberbullying suggest that this relationship needs to be further addressed. “Doing gender” varies across time and context; therefore, by considering different ways in which “doing gender” may vary will help scholars provide a more nuanced understanding regarding the relationship between gender and cyberbullying.

Intersectionality, Bullying, and Cyberbullying

Crenshaw’s (1989) *intersectionality* is a theoretical framework that allows researchers to consider experiences of multiple dimensions of individuals’ lives. An intersectional approach emphasizes that experiences with victimization may vary based on individuals’ social identities. That is, instead of looking at gender, racial, and sexuality-based categories as having separate and distinct effects on individuals’ experiences, an intersectional approach stresses that individuals’ social identities are experienced simultaneously (Choo & Ferree, 2010; Miller, 2006; Núñez, 2014). For example, Miller’s (2006) research on young African American women’s experiences with victimization both within and external to the school helped place a dually marginalized group at the center of research. Miller’s approach gave a voice to a marginalized group, whose experiences with victimization were largely overlooked.

Recently, Radu (2018b) argued that relatively few studies on bullying or cyberbullying examine the combined effects of gender, race, and sexuality on forms of bullying and outcomes associated with bullying. This problematic because youth who face intersecting categories of oppression, such as being a student of color and identifying as gay, lesbian, bisexual or transgender, may be at a higher risk for being targeted for both bullying and cyberbullying. In fact, a recent student survey administered to nearly 80,000 public school students found that two of the top three reasons students believe they are bullied are because of their race or skin color and because their peers perceive them as being gay (Youth Truth, 2016). This suggest that more research is needed that explicitly uses an intersectional approach to understand the bullying and cyberbullying victimization experiences of multiple-marginalized groups of youth.

IMPLICATIONS FOR POLICY AND DIRECTIONS FOR FUTURE RESEARCH

Most agree that schools are responsible for developing policies on both bullying and cyberbullying, including additional training for teachers and anti-cyberbullying initiatives (Macaulay, Betts, Stiller & Kellezi, 2018). Policies prohibiting the use of the school’s Internet system for cyberbullying behaviors are somewhat common across schools (Snakenborg, Van Acker, & Gable, 2011), and scholars continue to look for unique ways to recommend that schools address problems with cyberbullying. One example is Diamanduros and Downs (2011) who argued that within schools, the library media specialist is in a unique position to bring more awareness to students, faculty, and staff about ways to combat cyberbul-

lying. They argue that school library media specialists may be able to address cyberbullying in schools by establishing policies and promoting awareness to prevent online bullying.

Hoff and Mitchell (2009) agree that the school should be at the center of the solution to address youths' involvement with cyberbullying, but they recommend a more collaborative approach among educators, parents, and youths. They argue that first, students and their parents need to be educated about the dangers of cyberspace to protect youths from online victimization. However, Hoff and Mitchell also note that a limitation of this approach is it places most of the burden on those targeted by cyberbullying, rather than addressing the behaviors of the perpetrators. Moving forward, they recommend that it may be helpful to have students themselves engaged in the development of policies and practices to respond to cyberbullying because they may understand it better than school faculty, staff, and administrators.

Other studies also point to families as agents to help prevent cyberbullying. Recently, Radu, Sobba, and McManus (2017) argue that quality parent-child relationships and open communication may help deter bullying. Parents should both encourage their children to report bullying and online behaviors and understand what to do if they see bullying both in schools and online. Additionally, Beale and Hall (2007) suggests that in order to prevent cyberbullying, parents need to support schools' efforts to address cyberbullying, including talking to children about the appropriate use of the Internet, monitoring their children's Internet use, and staying up-to-date with trends in digital communication (e.g., acronyms and emojis). They also recommend that parents work in collaboration with teachers and law enforcement to create cyberbullying programs (Beale & Hall, 2007). Beale and Hall's ideas are supported by more recent studies that suggest that bridging families and schools may help deter forms of bullying. For example, Axford et al., (2015) suggest that parental involvement in school-based bullying programs may increase teachers' awareness of incidents of bullying through parental reporting. Consequently, teachers and/or other school administrators cannot address cyberbullying incidents if they are not made aware about the problematic behaviors occurring online.

While these studies point to directions for further development of bullying and cyberbullying prevention programs, the authors argue that families, schools, workplaces, and communities may be more successful in developing policies and programs to deter cyberbullying if they focus on gender differences associated with these problematic behaviors. For example, what could families and schools do to help youths construct positive masculinities that do not rely physically aggressive behaviors? What approaches would be useful for helping increasing girls' self-esteem, so they do not rely on putting their peers down in order to gain social status? What can school faculty, staff, and administrators do to build a more inclusive school environment for youths who do not "fit in" with their peers? How can workplaces develop information technology (IT) departments specifically designed to combat unethical and aggressive online behaviors? These questions point to directions for future research and policy development.

CONCLUSION

This article explored gender differences in bullying and cyberbullying using an interdisciplinary approach by merging research from a variety of disciplines. It also addressed commonalities and differences between bullying and cyberbullying and suggested how to utilize theory to better understand gender differences associated with cyberbullying. This was followed by a review of how cooperative forces among individuals, communities, and social institutions may be able deter both bullying and cyberbullying behaviors. Following current research, the authors argue that collaborative efforts among schools, families, workplaces, and communities may be most effective for deterring bullying that occurs within

various contexts and online. Additionally, considering how individuals “do gender” and how gender intersects with individuals’ other social identities may help in the development of future policies and programs that address problematic behaviors among both youths and adults.

REFERENCES

- Abreu, R. L., & Kenny, M. C. (2018). Cyberbullying and LGBTQ youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Child & Adolescent Trauma*, 11(1), 81–97. doi:10.1007/40653-017-0175-7 PMID:32318140
- Aivazpour, Z., & Beebe, N. (2018). *Cyberbullying: Investigating the roles of power and communication medium*. Association for Information Systems. Retrieved from <https://aisel.aisnet.org/amcis2018/Security/Presentations/35/>
- Axford, N., Farrington, D. P., Clarkson, S., Bjornstad, G. J., Wrigley, Z., & Hutchings, J. (2015). Involving parents in school-based programmes to prevent and reduce bullying: What effect does it have? *Journal of Children's Services*, 10(3), 242–251. doi:10.1108/JCS-05-2015-0019
- Balakrishnan, V. (2018). Actions, emotional reactions and cyberbullying—From the lens of bullies, victims, bully-victims and bystanders among Malaysian young adults. *Telematics and Informatics*, 35(5), 1190–1200. doi:10.1016/j.tele.2018.02.002
- Barlett, C. P., Gentile, D. A., Chng, G., Li, D., & Chamberlin, K. (2018). Social media use and cyberbullying perpetration: a longitudinal analysis. *Violence and Gender*, 5(3), 191–197.
- Carbone-Lopez, K., Esbensen, F. A., & Brick, B. T. (2010). Correlates and consequences of peer victimization: Gender differences in direct and indirect forms of bullying. *Youth Violence and Juvenile Justice*, 8(4), 332–350. doi:10.1177/1541204010362954
- Carter, J. M., & Wilson, F. L. (2015). Cyberbullying: A 21st century health care phenomenon. *Pediatric Nursing*, 41(3), 115–125. PMID:26201169
- Choo, H. Y., & Ferree, M. M. (2010). Practicing intersectionality in sociological research: A critical analysis of inclusions, interactions, and institutions in the study of inequalities. *Sociological Theory*, 28(2), 129–149. doi:10.1111/j.1467-9558.2010.01370.x
- Connell, R. W. (2005). *Masculinities*. University of California Press.
- Cooper, L. A., & Nickerson, A. B. (2013). Parent retrospective recollections of bullying and current views, concerns, and strategies to cope with children's bullying. *Journal of Child and Family Studies*, 22(4), 526–540. doi:10.1007/10826-012-9606-0
- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *University of Chicago Legal Forum*, 14, 538–554.
- Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development*, 66(3), 710–722. doi:10.2307/1131945 PMID:7789197

Cullerton-Sen, C., & Crick, N. R. (2005). Understanding the effects of physical and relational victimization: The utility of multiple perspectives in predicting social-emotional adjustment. *School Psychology Review, 34*(2), 147–160.

D'Souza, N., Forsyth, D., Tappin, D., & Catley, B. (2018). Conceptualizing workplace cyberbullying: Toward a definition for research and practice in nursing. *Journal of Nursing Management, 26*(7), 842–850. doi:10.1111/jonm.12614 PMID:29446175

Day, K. R., & Kahle, L. L. (2014). Gender, bullying victimization, and education. *Violence and Victims, 29*(5), 843–856. doi:10.1891/0886-6708.VV-D-13-00047 PMID:25905132

Diamanduros, T., & Downs, E. (2011). Creating a safe school environment: How to prevent cyberbullying at your school. *Library Media Connection, 30*(2), 36–38.

Festl, R., & Quandt, T. (2016). The role of online communication in long-term cyberbullying involvement among girls and boys. *Journal of Youth and Adolescence, 45*(9), 1931–1945. doi:10.1007/10964-016-0552-9 PMID:27474042

Goddard, C. (2008). H8@ skul: Cyber world bullying. *Education Digest, 73*(7), 4–9.

Goffman, E. (1976). *Gender display. Gender Advertisements. Communications and Culture*. Palgrave. doi:10.1007/978-1-349-16079-2

Gofin, R., & Avitzour, M. (2012). Traditional versus internet bullying in junior high school students. *Maternal and Child Health Journal, 16*(8), 1625–1635. doi:10.1007/10995-012-0989-8 PMID:22447501

Görzig, A. (2016). Adolescents' experience of offline and online risks: Separate and joint propensities. *Computers in Human Behavior, 56*, 9–13. doi:10.1016/j.chb.2015.11.006

Hoff, D. L., & Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration, 47*(5), 652–665. doi:10.1108/09578230910981107

Hutzell, K. L., & Payne, A. A. (2012). The impact of bullying victimization on school avoidance. *Youth Violence and Juvenile Justice, 10*(4), 370–385. doi:10.1177/1541204012438926

Konishi, C., Hymel, S., Zumbo, B. D., & Li, Z. (2010). Do school bullying and student—Teacher relationships matter for academic achievement? A multilevel analysis. *Canadian Journal of School Psychology, 25*(1), 19–39. doi:10.1177/0829573509357550

Kowalski, R. M., Toth, A., & Morgan, M. (2018). Bullying and cyberbullying in adulthood and the workplace. *The Journal of Social Psychology, 158*(1), 64–81. doi:10.1080/00224545.2017.1302402 PMID:28402201

Li, Q. (2006). Cyberbullying in schools: A research of gender differences. *School Psychology International, 27*(2), 157–170. doi:10.1177/0143034306064547

Macaulay, P. J., Betts, L. R., Stiller, J., & Kellezi, B. (2018). Perceptions and responses towards cyberbullying: A systematic review of teachers in the education system. *Aggression and Violent Behavior, 43*, 1–12. doi:10.1016/j.avb.2018.08.004

Mendez, J. J., Bauman, S., & Guillory, R. M. (2012). Bullying of Mexican immigrant students by Mexican American students: An examination of intracultural bullying. *Hispanic Journal of Behavioral Sciences*, 34(2), 279–304. doi:10.1177/0739986311435970

Messerschmidt, J. W. (2000). *Nine lives: Adolescent masculinities, the body, and violence*. Westview Press.

Miller, J. (2006). *Getting played: African American girls, urban inequality, and gendered violence*. New York University Press.

Motswi, E. M., & Mashegoane, S. (2017). The role of sex in the prevalence rates and psychological consequences of cyberbullying among learners in the Limpopo Province, South Africa. *Gender & Behaviour*, 15(3), 9808–9817.

Nansel, T. R., Craig, W., Overpeck, M. D., Saluja, G., & Ruan, W. J. (2004). Cross-national consistency in the relationship between bullying behaviors and psychosocial adjustment. *Archives of Pediatrics & Adolescent Medicine*, 158(8), 730–736. doi:10.1001/archpedi.158.8.730 PMID:15289243

Nishina, A., Juvonen, J., & Witkow, M. R. (2005). Sticks and stones may break my bones, but names will make me feel sick: The psychosocial, somatic, and scholastic consequences of peer harassment. *Journal of Clinical Child and Adolescent Psychology*, 34(1), 37–48. doi:10.120715374424jccp3401_4 PMID:15677279

Núñez, A. M. (2014). Employing multilevel intersectionality in educational research: Latino identities, contexts, and college access. *Educational Researcher*, 43(2), 85–92. doi:10.3102/0013189X14522320

Olweus, D. (1991). Bully/victim problems among schoolchildren: Basic facts and effects of a school-based intervention program. In D. J. Pepler & K. H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 411–448). Lawrence Erlbaum Associates.

Radu, M. B. (2018a). Do students' perceptions of unsafe schools and experiences with bullying hinder the effects of family and school social capital in deterring violence? *The American Behavioral Scientist*, 62(11), 1505–1524. doi:10.1177/0002764218787004

Radu, M. B. (2018b). Educational disparities among marginalized groups of students: how bully victimization and perceived lack of schools' safety impede students' educational attainment. In M. C. Grant (Ed.), *Equity, Equality, and Reform in Contemporary Public Education* (pp. 126–152). IGI Global. doi:10.4018/978-1-5225-4960-4.ch007

Radu, M. B., Sobba, K.N., & McManus, L. (2017). Who and what helps bully victims feel safe at school? *Sociation Today*, 15(2), Article 3.

Seals, D., & Young, J. (2003). Bullying and victimization: Prevalence and relationship to gender, grade level, ethnicity, self-esteem, and depression. *Adolescence*, 38(152), 735–747. PMID:15053498

Shultz, E., Heilman, R., & Hart, K. J. (2014). Cyber-bullying: An exploration of bystander behavior and motivation. *Cyberpsychology (Brno)*, 8(4), 3. doi:10.5817/CP2014-4-3

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, 49(2), 147–154.

Snakenborg, J., Van Acker, R., & Gable, R. A. (2011). Cyberbullying: Prevention and intervention to protect our children and youth. *Preventing School Failure*, 55(2), 88–95. doi:10.1080/1045988X.2011.539454

SooHoo, S. (2009). Examining the Invisibility of Girl-to-Girl Bullying in Schools: A Call to Action. *International Electronic Journal for Leadership in Learning*, 13(6).

Sticca, F., & Perren, S. (2013). Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*, 42(5), 739–750. doi:10.1007/10964-012-9867-3 PMID:23184483

Sullivan, T. N., Farrell, A. D., & Klierer, W. (2006). Peer victimization in early adolescence: Association between physical and relational victimization and drug use, aggression, and delinquent behaviors among urban middle school students. *Development and Psychopathology*, 18(1), 119–137. doi:10.1017/S095457940606007X PMID:16478555

Thomas, H. J., Connor, J. P., & Scott, J. G. (2015). Integrating traditional bullying and cyberbullying: Challenges of definition and measurement in adolescents—a review. *Educational Psychology Review*, 27(1), 135–152. doi:10.1007/10648-014-9261-7

Vranjes, I., Baillien, E., Vandebosch, H., Erreygers, S., & De Witte, H. (2018). Kicking someone in cyberspace when they are down: Testing the role of stressor evoked emotions on exposure to workplace cyberbullying. *Work and Stress*, 1–21.

Wang, J., Iannotti, R. J., & Nansel, T. R. (2009). School bullying among adolescents in the United States: Physical, verbal, relational, and cyber. *The Journal of Adolescent Health*, 45(4), 368–375. doi:10.1016/j.jadohealth.2009.03.021 PMID:19766941

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society*, 1(2), 125–151. doi:10.1177/0891243287001002002

Wilton, C., & Campbell, M. A. (2011). An exploration of the reasons why adolescents engage in traditional and cyber bullying. *Journal of Educational Sciences & Psychology*, 1(2), 101–109.

Wong, R. Y., Cheung, C. M., & Xiao, B. (2018). Does gender matter in cyberbullying perpetration? An empirical investigation. *Computers in Human Behavior*, 79, 247–257. doi:10.1016/j.chb.2017.10.022

Youth Truth. (2016). *Learning from student voice: How do student experience bullying?* San Francisco, CA: Youth Truth Student Survey: A National Nonprofit. Retrieved from <http://www.youthtruthsurvey.org/wp-content/uploads/2016/09/YouthTruth-Learning-from-Student-Voice-Bullying.pdf>

Zych, I., Baldry, A. C., Farrington, D. P., & Llorent, V. J. (2018). Are children involved in cyberbullying low on empathy? A systematic review and meta-analysis of research on empathy versus different cyberbullying roles. *Aggression and Violent Behavior*.

ADDITIONAL READINGS

Baldry, A. C., Farrington, D. P., & Sorrentino, A. (2017). School bullying and cyberbullying among boys and girls: Roles and overlap. *Journal of Aggression, Maltreatment & Trauma*, 26(9), 937–951. doi:10.1080/10926771.2017.1330793

Campbell, M., & Bauman, S. (Eds.). (2018). *Reducing cyberbullying in schools: International evidence-based best practices*. Academic Press.

Carvalho, M., Branquinho, C., & de Matos, M. G. (2018). Emotional symptoms and risk behaviors in adolescents: Relationships with cyberbullying and implications on well-being. *Violence and Victims*, 33(5), 871–885. doi:10.1891/0886-6708.VV-D-16-00204 PMID:30567870

DeSmet, A., Rodelli, M., Walrave, M., Soenens, B., Cardon, G., & De Bourdeaudhuij, I. (2018). Cyberbullying and traditional bullying involvement among heterosexual and non-heterosexual adolescents, and their associations with age and gender. *Computers in Human Behavior*, 83, 254–261. doi:10.1016/j.chb.2018.02.010

Lapidot-Lefler, N., & Dolev-Cohen, M. (2015). Comparing cyberbullying and school bullying among school students: Prevalence, gender, and grade level differences. *Social Psychology of Education*, 18(1), 1–16. doi:10.1007/11218-014-9280-8

Martínez, I., Murgui, S., Garcia, O. F., & Garcia, F. (2019). Parenting in the digital era: Protective and risk parenting styles for traditional bullying and cyberbullying victimization. *Computers in Human Behavior*, 90, 84–92. doi:10.1016/j.chb.2018.08.036

Olweus, D., & Limber, S. P. (2018). Some problems with cyberbullying research. *Current Opinion in Psychology*, 19, 139–143. doi:10.1016/j.copsyc.2017.04.012 PMID:29279213

Patchin, J. W., & Hinduja, S. (2010). Cyberbullying and self-esteem. *The Journal of School Health*, 80(12), 614–621. doi:10.1111/j.1746-1561.2010.00548.x PMID:21087257

Vranjes, I., Baillien, E., Vandebosch, H., Erreygers, S., & De Witte, H. (2017). The dark side of working online: Towards a definition and an Emotion Reaction model of workplace cyberbullying. *Computers in Human Behavior*, 69, 324–334. doi:10.1016/j.chb.2016.12.055

KEY TERMS AND DEFINITIONS

Bully: An individual who asserts dominance by engaging in repeated and aggressive behaviors towards an individual or individual(s).

Bully Victim: An individual who has been the target of harmful and aggressive behaviors.

Bullying: Intentional, repeated actions meant to harm a victim and perpetuate an inequity of power between the bullying and victim(s).

Cyberbullying: Aggressive behaviors that are perpetrated intentionally through mediated communication, such as email, text messages, social media, or other online websites.

Doing Gender: A perspective that suggests that gender is a social construct, a routine accomplishment that is embedded in everyday interactions.

Gender Display: The behavioral aspects of being a woman or man, opposed to biological differences between the sexes.

Intersectionality: A theoretical framework that considers multiple dimensions of individuals' lives, emphasizing that experiences may vary based on the overlapping nature of individuals' social identities.

Workplace Cyberbullying: Aggressive online behavior that is intentionally perpetrated between or among co-workers.

The Human Rights–Based Approach to Combat Cyberbullying Against Women and Girls

Sabrina Ching Yuen Luk

Nanyang Technological University, Singapore

INTRODUCTION

Cyberbullying against women and girls has become more prevalent when the Internet and mobile technologies become more widely available and accessible in terms of cost. The use of information and communications technologies (ICTs) by an individual or group to intentionally embarrass, taunt, harass, threaten, torment or humiliate a woman or a girl takes places in blogs, on chat rooms or social networking sites, through emails, voice mails, texting, instant messaging, photo or video-clip sharing. This has led to multifaceted consequences that negatively affect the woman's or the girl's self-esteem, physical and mental health, behaviours, relationship with other people and motivation to live, study and work. At present, actions taken to combat cyberbullying against women and girls remain inadequate because of insufficient attention and support and deeply held prejudices against women. In order to have an in-depth investigation into this issue, this study examines what cyberbullying is, its causes and consequences, and a human rights based approach to combat cyberbullying against women and girls.

BACKGROUND

Women and girls have long been victims of physical violence and sexual violence. According to the 2013 Global Estimates published by World Health Organization, about 1 in 3 (35 percent) of women worldwide “have experienced either physical and/or sexual intimate partner violence or non-partner sexual violence in their lifetime” (World Health Organization, 2017). Violence against women and girls “is prevalent in every society in the world” (UNiTE, n.d.) that it “cuts across boundaries of wealth, race and culture” (UNiTE, n.d.). It “is rooted in gender-based discrimination and social norms and gender stereotypes that perpetuate such violence” (UN Women, n.d.). Although some measures have been taken by governments across countries to end such violence, the continued prevalence of violence against women and girls is “yet to be tackled with all the necessary political commitment, action and resources” (UNiTE, n.d.).

Over the past decade, cyberbullying has become a new form of violence against women and girls. Cyberbullying refers to the use of information and communications technologies (ICTs) by an individual or group to intentionally embarrass, taunt, harass, threaten, torment or humiliate a person through a range of hostile and aggressive behaviours, such as trolling, stalking or doxing. Trolling disturbs or harms another person by posting irrelevant, unpleasant and offensive messages online while stalking causes a person to fear for his or her safety or suffer other emotional distress by willfully tracing, following and monitoring

DOI: 10.4018/978-1-7998-3479-3.ch029

such person online without his or her consent. Doxing involves revealing identifiable information about someone online without his or her consent, including one's real name, home address, phone number and email address (Amnesty International, 2018, p.29). Over the past decade, cyberbullying has become more prevalent when the Internet and mobile technologies become more widely available and accessible in terms of cost. It takes places in blogs, on chat rooms or social networking sites such as Facebook and Twitter. It also takes places through emails, voice mails, texting, instant messaging, photo or video-clip sharing. "As the bullying process continues, the frequency and intensity of the aggressions tend to increase" (Forssell, 2018). Cyberbullying involves an imbalance of power between the perpetrator and the victim due to the perpetrator having the expert knowledge of the Internet (Ybarra and Mitchell, 2004), the anonymity of the perpetrator, the aggression directed towards the victim and the fear created in the victim (Betts, 2016, pp.15-16). Cyberbullies are able to "hide their identities by using screen names and well-hidden internet protocol addresses, leaving the target vulnerable and unsettled" (Hoff and Mitchell, 2009, p.653). While most of the time the identities of cyberbullies are unknown to the victims, cyberbullies can be someone from the victim's social circle, such as a close friend, a classmate, a co-worker, a current or former boyfriend or girlfriend. In the virtual world, the removal of geographical and time constraints allows the acts associated with cyberbullying to be witnessed by an unlimited audience (Betts, 2016, p.39). This has the effect of the victim "experiencing repeated humiliation because although the act is not repeated the negative experience associated with it is" (Betts, 2016, p.38).

In recent years, cyberbullying has become alarmingly common among women and girls. A study that examined cyberbullying risk among teenagers in the United States (U.S.) found that adolescent females were more at risk for cyberbullying than adolescent males (Navarro and Jasinski, 2012, pp.89-90). A study that examined cyberbullying among 1,928 German, Dutch and Thai adolescents found that girls were more likely than boys to be victims while boys were more likely than girls to be bullies (Wachs et al., 2015). Another study that examined cyberbullying among 6,260 students aged 11-23 from six European countries found that girls were victims more often while boys were perpetrators more often (Schultze-Krumbholz et al., 2015). Similar results can also be found in Canada (Li, 2006; Ronis and Slaunwhite, 2017), Sweden (Låftman et al, 2013), Israel (Heiman and Olenik-Shemesh, 2015), Belfast (Savoldi and Ferraz de Abreu, 2016) and Nigeria (Olumide et al., 2016). A study that examined over 5,300 students aged 11-15 in the United Kingdom (UK) found that girls were twice as likely as boys to report being cyberbullied (Public Health England, 2014, p.4). A study that examined cyberbullying among university students in Canada found that more female than male students experienced cyberbullying by a friend or acquaintance at the university and "female students were much more likely than males to report having experienced cyberbullying over social networks and via text messages" (Faucher et al., 2014). A large-scale study on cyberbullying among 31,148 teenagers found that cyberbullying victimization was more prevalent in females than males and female adolescents were more vulnerable to emotional and behavioural problems as a result of cyberbullying (Kim et al., 2018, p.669). In 2017, a survey conducted by Amnesty International revealed that about a quarter of 4,000 women surveyed across eight countries had experienced online abuse or harassment at least once and over 40 percent of them said that such experience threatened their physical safety (Amnesty International, 2017). The problem of cyberbullying against women and girls is a serious problem that deserves an in-depth investigation.

THE CAUSES OF CYBERBULLYING AGAINST WOMEN AND GIRLS

There are some reasons attributed to cyberbullying against women and girls. Female students are victims of cyberbullying because of their physical appearance, having different opinions or beliefs, upsetting other people, interpersonal problems (Faucher et al., 2014) and learning disabilities (Heiman and Olenik-Shemesh, 2015). They are also victims of cyberbullying due to relationship problems, including break-ups, envy, intolerance, and ganging up (Hoff and Mitchell, 2009, p.655). Break-ups that cause feeling of rejection lead to retaliation by ex-partner to publicly shame and humiliate the victim for ending a relationship (European Institute for Gender Equality, 2017; Hoff and Mitchell, 2009). Being envious of someone's achievements or popularity or being intolerant of someone's religion, disability and gender make the bully resort to online channels to vent anger and frustration (Hoff and Mitchell, 2009). Cyberbullying is "often due to ganging up on a person in order to reject and isolate that person from the 'group'" (Hoff and Mitchell, 2009, p.656). Cyberbullying can be seen as a continuum of bullying in the physical space. Female students who are victims of traditional bullying are also more likely to be victims of cyberbullying (Smith et al., 2008; Vandebosch and Van Cleemput, 2009).

It is common for females being targeted for cyberbullying due to their refusal to reciprocate the bully's romantic interest in them (D'Cruz & Noronha, 2013). Apart from being physically stalked by the bully, females are sexually harassed by the bully through emails (D'Cruz & Noronha, 2013, p.331). Their office computer systems, email and social media site accounts are hacked by the bully and their email identities are used by the bully to contact their family and friends (D'Cruz & Noronha, 2013). Cyberbullying can be committed by ex-partner or partners to seek interpersonal revenge or entertainment through distributing or sharing texts, non-consensual private or sexual images or videos of females online such as revenge porn-specific sites (Hall and Hearn, 2017). It is another form of gender violence and abuse representing a perception of regaining control and reaffirming a sense of masculinity (Hall and Hearn, 2017).

Workplace cyberbullying against women can come from their bosses, peers, subordinates or even former employees. This is due to the boss's "personality traits especially the needs to be in control and to get even in instances of perceived defiance" (D'Cruz & Noronha, 2013, p.332), peers being threatened by the female's competence, subordinates' grievance that were unable to share directly with their superiors, and former employees' revenge for being laid off from work (D'Cruz & Noronha, 2013). Low support from managers was related to higher exposure to cyberbullying behaviours for women non-managers while low support from colleagues was associated with exposure to cyberbullying behaviours for women managers (Forssell, 2018). Women also face online violence and abuse due to speaking out about some feminist issues, standing up to injustice and being public figures (Amnesty International, 2018, pp.14-8). Prominent women leaders, journalists, rights activists and bloggers are usually subjected to violence and abuse online, "especially when it is related to fields where men have been traditionally held as experts, such as gaming, politics and technology" (Amnesty International, 2018, p.15). In order to silence and exclude the voices of these women, the bully send violent and threatening messages such as death threats or rape threats to them online (Amnesty International, 2018). Cyber violence is used as a way of disciplining women for transgressions of traditional gender boundaries and force them into a subordinate position compared with men. Women also face violence and abuse online due to racial, religious, disability discrimination or discrimination based on sexual orientation or gender identity (Amnesty International, 2018).

THE CONSEQUENCES OF CYBERBULLYING AGAINST WOMEN AND GIRLS

3

Different studies show that cyberbullying against women and girls can lead to multifaceted consequences that negatively affect victims' self-esteem, physical and mental health, behaviours, relationship with other people and motivation to live, study and work. Female secondary school students who were victims of cyberbullying had lower body esteem that led to lower appearance and weight satisfaction (Olenik-Shemesh and Heiman, 2017). They had poorer subjective health, including headache, upset stomach, poor appetites, and sleep disturbances (Låftman et al., 2013). They had lower perceived peer acceptance, value of learning and importance of school (Betts et al., 2017). They had greater absenteeism and poorer grades in school. They felt helpless and lonely, had disillusionment and distrust of other people, social isolation and were reluctant to meet new people (Šléglová and Cerna, 2011). They showed signs of physical aggression targeted at other people (Šléglová and Cerna, 2011) and had more behavioral problems such as physical violence and conduct disorder (Kim et al., 2018). They had lower life satisfaction (You et al., 2016) but higher rates of emotional problems such as depression, suicidal behaviors (Reed et al., 2015) and substance use problems than their male counterparts (Elgar et al., 2014). Female college students who had experienced cyberbullying had almost three times the odds of meeting clinical criteria for depression compared to those with no cyberbullying experience (Selkie et al., 2015, p.82). They lost confidence and self-esteem, felt embarrassed, angry, frustrated and afflicted due to the lack of control with the situation (Brewer et al., 2012). University female students who had been cyberbullied reported that cyberbullying threatened their emotional security or physical safety, affected both physical health (headaches, stomach problems, nausea, heart palpitations or chest pain, sweating, etc.) and mental health (e.g. anxiety, depression, emotional outbursts, etc.) (Faucher et al., 2014). Cyberbullying also affected their ability to do assignments (e.g. productivity, loss of confidence, concentration problems, etc.) and obtain better grades, their personal relationships at the university and outside of the university and made them want to cyberbully back (Faucher et al., 2014).

Women that experienced cyberbullying at work had lower psychological health (i.e. depression, anxiety, and psychosomatic health) (Ford, 2013), higher mental strain, lower job satisfaction (Coyne, 2017) and lower task performance (D'Cruz & Noronha, 2013). They felt fatigued (Sansone and Sansone, 2015), scared and powerless, had physical and emotional distress, emotional exhaustion, headaches, blood pressure and gastrointestinal problems (D'Cruz & Noronha, 2013). Those who had experienced sexual harassment online suffered great trauma with poor physical and mental conditions (D'Cruz & Noronha, 2013). They were under great shock and stress and were unable to eat and sleep (D'Cruz & Noronha, 2013). They feared for their physical safety and worried about their personal and professional reputation (D'Cruz & Noronha, 2013, p.336). Cyberbullying can extend to the physical world when doxing can lead to stalking in the real world, undermining the physical safety of the victims and their family members. The survey conducted by Amnesty International revealed that cyberbullying had devastating psychological impact on female victims, with 63 percent of respondents having insomnia, 60 percent experiencing lower self-esteem or loss of self-confidence, 56 percent being unable to concentrate for long periods of time, 55 percent experiencing stress, anxiety or panic attacks and 24 percent feared for their family's safety (Amnesty International, 2017). The survey also showed that cyberbullying undermined female victims' freedom of expression, leading to victims refraining from expressing their opinions online, stepping back from public conversations or self-censoring out of fear for their safety or privacy (Amnesty International, 2017). Some may even choose to leave social media platforms to stop receiving abuse (Dhrodia, 2018, p.385).

SOLUTIONS AND RECOMMENDATIONS

Cyberbullying against women and girls is a form of violence that is just as damaging to women as physical violence (African Development Bank Group, 2016). It is also a form of discrimination that reinforces prejudice and a major impediment to achieving gender equality and the full advancement of women because of severely violating the principle of respect for human dignity and a range of women's rights, including the rights to life, equality, privacy, freedom of opinion and expression, and to be free from any physical, sexual or psychological violence. In fact, women deserve equal rights as men, which is enshrined in the International Covenant on Civil and Political Rights, the International Covenant on Economic, Social and Cultural Rights, and the Convention on the Elimination of Discrimination Against Women (CEDAW). Besides, "[r]ights of women in cyber space are as important as rights of women in physical space" (Halder and Jaishankar, 2012, p.55). It is important to identify and protect women's right in the cyber space so that women can be free from discrimination and violence. The adoption of a human rights based approach to combat cyberbullying against women and girls is to encourage rights holders to claim their rights as well as developing the capacity of duty-bearers to meet their obligations (United Nations Population Fund, 2014). Women being rights holders have the rights to privacy and freedom of expression online and to be free from any physical, sexual or psychological violence. Meanwhile, the adoption of a human rights based approach recognizes that duty bearers have the obligations to respect, protect and guarantee these rights (United Nations Population Fund, 2014). Being duty bearers, the government, private organizations and non-government organizations (NGOs) should formulate and implement policies to combat cyberbullying against women and girls.

Women have the right to safe and protected cyber life that allows them to be free from all sorts of gender abuses and harassments (Halder and Jaishankar, 2012, p.62), especially their rights to privacy and freedom of expression. It is necessary to provide women and girls with equal treatment for legal protection against cyber abuses and harassments (Halder and Jaishankar, 2012) so that they can obtain the justice they deserve. At present, most of the new laws or amended laws that are introduced to combat cyberbullying by countries are gender neutral in nature. For example, in the U.S., 48 out of 50 states have included cyberbullying or online harassment in their bullying laws while 44 states have criminal sanction for cyberbullying or electronic harassment (Cyberbullying Research Center, n.d.). Depending on the severity of cyberbullying, the perpetrator can face fines or imprisonment. Besides, 45 states have included school sanction for cyberbullying (Cyberbullying Research Center, n.d.). For example, Public Act 478 in Michigan requires schools to modify their anti-bullying policy to include cyberbullying as a form of bullying (State of Michigan, 2015) so that they have the duty to intervene on any reports of cyberbullying. The Criminal Code Act 1995 in Australia also addresses cyberbullying. In India, the Information Technology Act, 2000 deals with offences such as hacking with computer systems and breaches of confidentiality and privacy. Section 509 of the Indian Penal Code, which makes word, gesture or act intended to insult the modesty of a woman punishable, can be applied to address sexual abuse faced by women online (Padte, 2012). Many governments have yet to strengthen the legal system to protect women's rights against cyberbullying. As Article 2 of CEDAW states that state parties agree to "adopt appropriate legislative and other measures, including sanctions where appropriate, prohibiting all discrimination against women" (United Nations, n.d.), governments can consider drafting and introducing gender-centric law on cyberbullying based on the principles set forth in CEDAW.

Legislation alone is not enough to combat cyberbullying against women and girls. Other mechanisms should be provided by the government for victims and witnesses to report suspected cases of cyberbullying against women. For example, the Child Exploitation and Online Protection Centre (CEOP), which is

a command of National Crime Agency established in 2006 in the UK, allows the public to report cases of online sexual abuse and exploitation of children on its website. Up until September 2009, the CEOP had safeguard 515 children, arrested 821 suspected offenders, disrupted and dismantled 205 high risk sex offender networks (The Secretary of State for the Home Department, 2010, p.9). The Australian Cybercrime Online Reporting Network (ACORN) is a national online system that allows the public to securely report instances of cyberbullying and cybercrimes such as hacking and identity theft as well as providing advice for victims (Australian Cybercrime Online Reporting Network, n.d.). It does not have any investigative or recovery powers, but has been delivered in collaboration with all Australian police agencies, Attorney-General's Department, Criminal Intelligence Commission and so on (Australian Cybercrime Online Reporting Network, n.d.).

Another equally important issue is to provide sufficient training to law enforcement officers so that they can have the capacity to proactively prevent, prepare for and respond to cyberbullying incidents (Olsen, 2014). They should have a good understanding of laws related to cyberbullying (Olsen, 2014) and “maintain a working knowledge of the latest technology and social networking sites to successfully investigate cyberbullying cases” (Thaxter, 2010, p.530). They should work with the communities such as schools to develop clear procedures for determining if and how officers should respond to cyberbullying complaints (Olsen, 2014). They can form partnership with schools through a written memorandum of understanding to addresses the specific issue of cyber bullying, such as the issue of jurisdiction (Thaxter, 2010, p.530). They should know how to collect and document evidence of cyberbullying incidents, communicate with the victims and their family members, provide support to the victims, and increase public awareness of cyberbullying behaviors through education workshops (Olsen, 2014). For example, juvenile police officers in Massachusetts adopt a three-pronged strategy, namely education, intervention and adjudication, to handle cyberbullying incidents in schools (Thaxter, 2010, p.529).

Meanwhile, the government and NGOs should provide support services such as counselling service, referral service, mental health services and legal aids for women and girls subject to cyberbullying. For example, a UK-based charity Family Lives provides a confidential helpline for young people subject to cyberbullying to seek support and advice. Help123 in Singapore is a one-stop service for youth, family members and educators to get updated cyber wellness resources and appropriate support services from cyber wellness community agencies through hotline, email or webchat (Fei Yue Community Services and TOUCH Cyber Wellness, 2017). In Australia, the Office of the eSafety Commissioner provides channels for the public to report cyberbullying incidents, counselling services for those in need and online safety information and education resources for young people, parents, women and adults. Local Community Legal Centres in Australia allows those who are subject to cyber abuse to seek free legal advice (The Office of the eSafety Commissioner, n.d.). Besides, the government and NGOs should implement anti-cyberbullying campaigns to eliminate stereotypes about gender roles and sexual norms for women, raise public awareness of cyberbullying and empower the public with the know-how to prevent cyberbullying. For example, Better Internet Campaign was launched in Singapore to combat cyberbullying and raise awareness of cyberwellness issues through a series of outreach events in secondary schools and tertiary institutions (Hio, 2017 February 7). The campaign website contained personal stories and experiences of cyberbullying, which were told from the perspectives of bully and victim (Hio, 2017 February 7).

Some social media sites are criticized for their inconsistency and inaction on their own rules when women report violence and abuse (Amnesty International, 2018, p.33). They are criticized for taking action only when they are under pressure (Amnesty International, 2018, p.40) and “being considerably more responsive to reports which appear in the media, including those highlighted in newspaper exposés” (Home Affairs Committee, 2017, p.15). Providers of social media sites should perform due diligence

by developing effective policies that act against cyberbullying against women on their platforms. They should publish clear community guidelines to prohibit online violence and abuse against women. Besides, they have safeguarding responsibilities to proactively and quickly identify and remove abusive and discriminatory online content such as hate speech against women and unwanted sexual advances. They should lock or suspend the accounts of those who violate community guidelines and provide tools for women to protect themselves from cyberbullying by blocking cyberbullies, muting conversations and filtering out obscene contents or offensive remarks. They should provide a report link for users to report potentially breaching content and be more responsive to cyberbullying complaints. They should increase transparency by regularly publishing reports on their efforts to analyse and handle complaints of cyberbullying against women. They should “create appropriate record keeping systems specific to violence against women and classify and share the ways in which they have responded to it” (Amnesty International, 2018, p.45).

Employers have the responsibility to provide employees with a safe and healthy working environment free of cyberbullying behaviors. They should “adopt a clear policy, along with reporting channels and processes, to indicate to employees that the organisation takes cyberbullying seriously” (D’Souza, 2018). They should formulate policies that allow early intervention to de-escalate issues of cyberbullying and carry out disciplinary action for employees violation of workplace bullying policy. Besides, they should implement workplace bullying prevention program that also incorporates anti-cyberbullying practice. Preventive measures such as the introduction of web-filtering system and the provision of routine in-service training programs on respectful workplace interaction online should be implemented.

FUTURE RESEARCH DIRECTIONS

There is a pressing need to combat cyberbullying against women and girls at the national and international levels. However, actions taken to combat cyberbullying against women and girls are still inadequate. A human right based approach to combat cyberbullying against women and girls has not been strongly advocated in different countries. This may be due to the fact that stereotypes about gender roles and sexual norms for women are still deeply rooted in today’s society. Future research directions can examine whether the adoption a human right based approach can effectively combat cyberbullying against women and girls and examine best practices in combating cyberbullying against women and girls.

CONCLUSION

In conclusion, cyberbullying against women and girls is a major impediment to achieving gender equality and the full advancement of women. The adoption of a human rights based approach to combat cyberbullying against women and girls is to encourage women to claim their rights to privacy and freedom of expression online and to be free from any physical, sexual or psychological violence while recognizing the obligations of governments, employers, NGOs and social media site providers to respect, protect and guarantee these rights. Combatting cyberbullying against women and girls is an ongoing process that requires efforts to eliminate prejudice and the determination and commitment to create a culture of respect in the cyber space. Looking forward, it is anticipated that women and girls can enjoy the same rights as men when they are online.

REFERENCES

- African Development Bank Group. (2016). *Minding the gaps: Identifying strategies to address gender-based cyber violence in Kenya*. Retrieved May 12, 2018, from https://www.afdb.org/fileadmin/uploads/afdb/Documents/Generic-Documents/Policy_Brief_on_Gender-Based_Cyber_Violence_in_Kenya.pdf
- Amnesty International. (2017). *Amnesty reveals alarming impact of online abuse against women*. Retrieved May 11, 2018, from <https://www.amnestyusa.org/press-releases/amnesty-reveals-alarming-impact-of-online-abuse-against-women/>
- Amnesty International. (2018). *#TOXICTWITTER: Violence and abuse against women online*. Retrieved May 11, 2018, from <https://www.amnestyusa.org/wp-content/uploads/2018/03/Toxic-Twitter.pdf>
- Australian Cybercrime Online Reporting Network. (n.d.). *Australian Cybercrime Online Reporting Network*. Retrieved May 13, 2018, from <https://www.acic.gov.au/our-services/australian-cybercrime-online-reporting-network>
- Betts, L. R. (2016). *Cyberbullying: Approaches, consequences and interventions*. London: Palgrave Macmillan; Palgrave Pivot. doi:10.1057/978-1-137-50009-0
- Betts, L. R., Spenser, K. A., & Gardner, S. E. (2017). Adolescents' involvement in cyber bullying and perceptions of school: The importance of perceived peer acceptance for female adolescents. *Sex Roles*, 77(7-8), 471–481. doi:10.1007/11199-017-0742-2 PMID:28979061
- Brewer, B., Cave, A., Massey, A., Vurdelja, A., & Freeman, J. (2012). Cyber bullying among female college students: An exploratory study. *Californian Journal of Health Promotion*, 12(1), 40–51. doi:10.32398/cjhp.v12i1.1554
- Coyne, I., Farley, S., Axtell, C., Sprigg, C., Best, L., & Kwok, O. (2017). Understanding the relationship between experiencing workplace cyberbullying, employee mental strain and job satisfaction: A dysempowerment approach. *International Journal of Human Resource Management*, 28(7), 945–972. doi:10.1080/09585192.2015.1116454
- Cyberbullying Research Center. (n.d.). *Bullying Laws Across America*. Retrieved May 13, 2018, from <https://cyberbullying.org/bullying-laws>
- D'Cruz, P., & Noronha, E. (2013). Navigating the extended reach: Target experiences of cyberbullying at work. *Information and Organization*, 23(4), 324–343. doi:10.1016/j.infoandorg.2013.09.001
- D'Souza, N. (2018). *What employers need to do to protect workers from cyberbullying*. Retrieved May 14, 2018, from <https://theconversation.com/what-employers-need-to-do-to-protect-workers-from-cyberbullying-87499>
- Dhrodia, A. (2018). Unsocial media: A toxic place for women. *IPPR Progressive Review*, 24(4), 381–387. doi:10.1111/newe.12078
- Elgar, F. J., Napoletano, A., Saul, G., Dirks, M. A., Craig, W., Poteat, V. P., Holt, M., & Koenig, B. W. (2014). Cyberbullying victimization and mental health in adolescents and the moderating role of family dinners. *JAMA Pediatrics*, 168(11), 1015–1022. doi:10.1001/jamapediatrics.2014.1223 PMID:25178884
- European Institute for Gender Equality. (2017). *Cyber violence against women and girls*. Retrieved May 13, 2018, from <https://eige.europa.eu/rdc/eige-publications/cyber-violence-against-women-and-girls>

- Faucher, C., Jackson, M., & Cassidy, W. (2014). Cyberbullying among university students: Gendered experiences, impacts, and perspectives. *Education Research International*, 2014, 698545. Retrieved May 9, 2018, from <https://www.hindawi.com/journals/edri/2014/698545/>. doi:10.1155/2014/698545
- Fei Yue Community Services and TOUCH Cyber Wellness. (2017). *About*. Retrieved May 13, 2018, from <https://help123.sg/about>
- Ford, D. P. (2013). Virtual harassment: Media characteristics' role in psychological health. *Journal of Managerial Psychology*, 28(4), 408–428. doi:10.1108/JMP-12-2012-0398
- Forssell, R. C. (2018). Gender and organisational position: Predicting victimisation of cyberbullying behaviour in working life. *International Journal of Human Resource Management*, ●●●, 1–20. doi:10.1080/09585192.2018.1424018
- Halder, D., & Jaishankar, K. (2012). *Cyber Crime and the Victimization of Women: Laws, Rights, and Regulations*. IGI Global. doi:10.4018/978-1-60960-830-9
- Hall, M., & Hearn, J. (2017). Revenge pornography and manhood acts: A discourse analysis of perpetrators' accounts. *Journal of Gender Studies*, 1–13. doi:10.1080/09589236.2017.1417117
- Heiman, T., & Olenik-Shemesh, D. (2015). Cyberbullying experience and gender differences among adolescents in different educational settings. *Journal of Learning Disabilities*, 48(2), 146–155. doi:10.1177/0022219413492855 PMID:23784784
- Hio, L. (2017 February 7). *Campaign for a Better Internet launched, with focus on cyberbullying, fake news*. Retrieved May 14, 2018, from <https://www.straitstimes.com/tech/campaign-for-a-better-internet-launched-with-focus-on-cyberbullying-fake-news>
- Hoff, D. L., & Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration*, 47(5), 652–665. doi:10.1108/09578230910981107
- Home Affairs Committee. (2017). *Hate crime: abuse, hate and extremism online*. Retrieved May 13, 2018, from <https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/609.pdf>
- Kim, S., Colwell, S. R., Kata, A., Boyle, M. H., & Georgiades, K. (2018). Cyberbullying victimization and adolescent mental health: Evidence of differential effects by sex and mental health problem type. *Journal of Youth and Adolescence*, 47(3), 661–672. doi:10.1007/10964-017-0678-4 PMID:28434092
- Låftman, S. B., Modin, B., & Östberg, V. (2013). Cyberbullying and subjective health: A large-scale study of students in Stockholm, Sweden. *Children and Youth Services Review*, 35(1), 112–119. doi:10.1016/j.childyouth.2012.10.020
- Li, Q. (2006). Cyberbullying in schools: A research of gender differences. *School Psychology International*, 27(2), 157–170. doi:10.1177/0143034306064547
- Navarro, J. N., & Jasinski, J. L. (2012). Going cyber: Using routine activities theory to predict cyberbullying experiences. *Sociological Spectrum*, 32(1), 81–94. doi:10.1080/02732173.2012.628560
- Olenik-Shemesh, D., & Heiman, T. (2017). Cyberbullying victimization in adolescents as related to body esteem, social support, and social self-efficacy. *The Journal of Genetic Psychology*, 178(1), 28–43. doi:10.1080/00221325.2016.1195331 PMID:27355106

Olsen, J. (2014). *Tips for law enforcement: Preparing for and responding to cyberbullying*. Retrieved May 13, 2018, from http://msue.anr.msu.edu/news/tips_for_law_enforcement_preparing_for_and_responding_to_cyberbullying

Olumide, A. O., Adebayo, E., & Oluwagbayela, B. (2016). Gender disparities in the experience, effects and reporting of electronic aggression among secondary school students in Nigeria. *BMJ Global Health*, 1(3), e000072. doi:10.1136/bmjgh-2016-000072 PMID:28588956

Padte, R. K. (2012). *Section 66A, Sexual Harassment and Women's Rights*. Retrieved January 3, 2018 from <https://internetdemocracy.in/2012/12/section-66a-sexual-harassment-and-womens-rights-2/>

Public Health England. (2014). *Cyberbullying: An analysis of data from the Health Behaviour in School-aged Children (HBSC) survey for England, 2014*. Retrieved May 9, 2018, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/621070/Health_behaviour_in_school_age_children_cyberbullying.pdf

Reed, K. P., Nugent, W., & Cooper, R. L. (2015). Testing a path model of relationships between gender, age, and bullying victimization and violent behavior, substance abuse, depression, suicidal ideation, and suicide attempts in adolescents. *Children and Youth Services Review*, 55, 128–137. doi:10.1016/j.childyouth.2015.05.016

Ronis, S., & Slaunwhite, A. (2017). *Gender and geographic predictors of cyberbullying victimization, perpetration, and coping modalities among youth*. Retrieved May 9, 2018, from <https://journals.sagepub.com/doi/abs/10.1177/0829573517734029>

Sansone, R. A., & Sansone, L. A. (2015). Workplace bullying: A tale of adverse consequences. *Innovations in Clinical Neuroscience*, 12(1-2), 32–37. <http://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=1&sid=ff4d8714-58bb-4dbb-b2db-8cfd629e6885%40pdv-v-sessmgr04> PMID:25852978

Savoldi, F., & Ferraz de Abreu, P. (2016). Bullying, cyberbullying and Internet usage among young people in post-conflict Belfast. *Cogent Social Sciences*, 2(1), 1132985. doi:10.1080/23311886.2015.1132985

Schultze-Krumbholz, A., Göbel, K., Scheithauer, H., Brighi, A., Guarini, A., Tsorbatzoudis, H., Barkoukis, V., Pyzalski, J., Plichta, P., Del Rey, R., Casas, J. A., Thompson, F., & Smith, P. K. (2015). A comparison of classification approaches for cyberbullying and traditional bullying using data from six European countries. *Journal of School Violence*, 14(1), 47–65. doi:10.1080/15388220.2014.961067

Selkie, E. M., Kota, R., Chan, Y., & Moreno, M. (2015). Cyberbullying, depression, and problem alcohol use in female college students: A multisite study. *Cyberpsychology, Behavior, and Social Networking*, 18(2), 79–86. doi:10.1089/cyber.2014.0371 PMID:25684608

Šléglová, V., & Cerna, A. (2011). Cyberbullying in adolescent victims: Perception and coping. *Cyberpsychology (Brno)*, 5(2), 4. Retrieved May 9, 2018, from <https://cyberpsychology.eu/article/view/4248/3294>

Smith, P. J., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 49(4), 376–385. doi:10.1111/j.1469-7610.2007.01846.x PMID:18363945

State of Michigan. (2015). *Michigan Public Act 478 of 2014 – Cyberbullying (MCL 380.1310b)*. Retrieved May 13, 2018, from https://www.michigan.gov/documents/mde/Cyber_Bullying_481850_7.pdf

Thaxter, K. (2010). Cyber bullying: Challenges and strategies faced by juvenile police officers. *Journal of Social Sciences*, 6(4), 529–531. <https://thescipub.com/pdf/10.3844/jssp.2010.529.531>. doi:10.3844/jssp.2010.529.531

The Office of the eSafety Commissioner. (n.d.). *Cyber abuse*. Retrieved May 13, 2018, from <https://www.esafety.gov.au/esafety-information/esafety-issues/cyber-abuse>

The Secretary of State for the Home Department. (2010). *Child Exploitation and Online Protection Centre (CEOP): The way forward*. Retrieved May 13, 2018, from <https://dera.ioe.ac.uk/10569/1/7785.pdf>

UN Women. (n.d.). *Focusing on prevention to stop the violence*. Retrieved August 1, 2018, from <https://www.unwomen.org/en/what-we-do/ending-violence-against-women/prevention>

UNiTE. (n.d.). *Human rights violation*. Retrieved August 1, 2018, from <https://www.un.org/en/women/endviolence/situation.shtml>

United Nations. (n.d.). *Convention on the Elimination of All Forms of Discrimination against Women*. Retrieved May 13, 2018, from <https://www.un.org/womenwatch/daw/cedaw/text/econvention.htm>

United Nations Population Fund. (2014). *The human rights-based approach*. Retrieved May 13, 2018, from <https://www.unfpa.org/human-rights-based-approach>

Vandebosch, H., & Van Cleemput, K. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society*, 11(8), 1349–1371. doi:10.1177/1461444809341263

Wachs, S., Junger, M., & Sittichai, R. (2015). Traditional, cyber and combined bullying roles: Differences in risky online and offline activities. *Societies (Basel, Switzerland)*, 5(1), 109–135. doi:10.3390/soci5010109

World Health Organization. (2017). *Violence against women*. Retrieved August 1, 2018, from <https://www.who.int/news-room/fact-sheets/detail/violence-against-women>

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 45(7), 1308–1316. doi:10.1111/j.1469-7610.2004.00328.x PMID:15335350

You, S., Lee, Y., & Kim, E. (2016). Physical, social, and cyberbullying: Relationships with adolescents' psychosocial factors. *Child Indicators Research*, 9(3), 805–823. doi:10.1007/12187-015-9338-y

ADDITIONAL READING

Choja, O., & Nelson, O. (2016). Psychological violence and the bane of cyber-harassment against women: An experiential inquest on Facebook. *Gender & Behaviour*, 14(3), 7589–7608.

Chukwuere, J. E., & Chukwuere, P. C. (2017). Cyber bullying of female students: An exploration of literature study. *Gender & Behaviour*, 15(4), 9983–9995.

Donner, C. M. (2016). The Gender Gap and Cybercrime: An Examination of College Students' Online Offending. *Victims & Offenders*, 11(4), 556–577. doi:10.1080/15564886.2016.1173157

Eckert, S. (2018). Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States. *New Media & Society*, 20(4), 1282–1302. doi:10.1177/1461444816688457

Fox, J., Cruz, C., & Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52, 436–442. doi:10.1016/j.chb.2015.06.024

Halder, D. (2015). Cyber stalking victimisation of women: Evaluating the effectiveness of current laws in India from restorative justice and therapeutic jurisprudential perspectives. *Temida: Casopis o Viktimizaciji, Ljudskim Pravima i Rodu*, 18(3-4), 103–130. doi:10.2298/TEM1504103H

Henry, N., & Powell, A. (2015). Embodied Harms: Gender, Shame, and Technology-Facilitated Sexual Violence. *Violence Against Women*, 21(6), 758–779. doi:10.1177/1077801215576581 PMID:25827609

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47(Part A), 46-55. doi:10.1016/j.wsif.2014.07.012

KEY TERMS AND DEFINITIONS

Cyberbullying: The use of information and communications technologies (ICTs) by an individual or group to intentionally embarrass, taunt, harass, threaten, torment, or humiliate a person.

Depression: A mental disorder that causes a person to have constant feelings of sadness and worthlessness and lose interests in study, work, friends, hobbies and so on.

Freedom of Expression: The right of a person to freely express himself or herself orally or in writing.

Gender Equality: Equal rights and opportunities enjoyed by men and women.

Right to Privacy: The right of a person to be free from public scrutiny or intrusion into matters of a personal nature.

Stalking: To willfully trace, follow and monitor a person online without his or her consent.

Trolling: To disturb or harm another person by posting irrelevant, unpleasant, and offensive messages online.

Section 4

Cyber Security and Data Protection

Hexa-Dimension Metric, Ethical Matrix, and Cybersecurity

4

Wanbil William Lee

The Computer Ethics Society, Hong Kong & Wanbil and Associates, Hong Kong

INTRODUCTION

Enormous spending on protecting data continues, hacking continues, yet data protection must helplessly go on. This is attributable to a *vicious circle* born out of a technology-driven information-intensive *modus operati*, culminating in a *chronic problem* which is complicated by a *tripartite relationship* and exacerbated by a muddled view of the key concepts which hinders a wholesome appreciation of the problem. The consequence amounts to business opportunities for the data protection enterprises and employment opportunities for the information security practitioners but a nightmare that haunts corporate and information security managers and the netizens at large. Given the status quo, an ethics-based framework was perceived and developed with the aim to mitigate rather than eradicate, to lessen the incidence of hacking or make hacking exasperate. Grounded in Ethical Computing (Lee, 2015b), this ethical approach aims to identify the concerns of the stakeholders against the ethical values, and to evaluate and balance the issues thus identified and discovered which invariably involve conflicting interests. This Ethical Matrix and Hexa-dimension Metric (Lee, 2019) poise to lend a hand. This article is an exposition of the problem that leads to the new methodology and will end with an illustration of the newly adapted/developed tools.

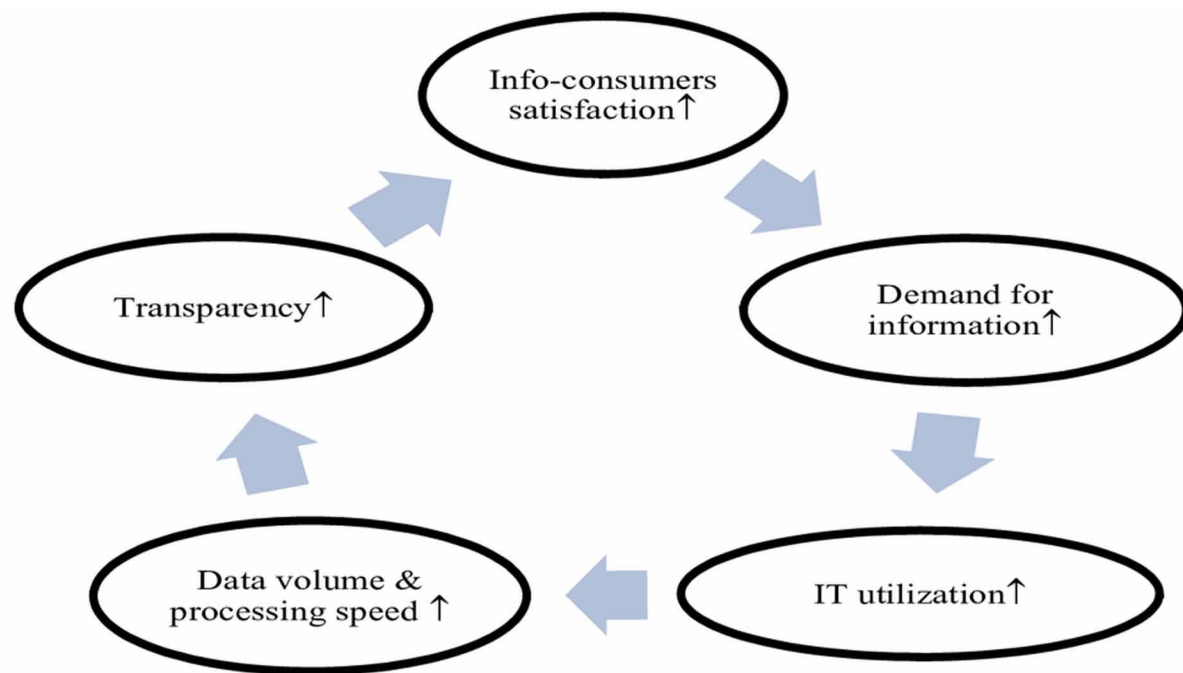
BACKGROUND

Data Generation/Consumption/Protection – A Vicious Circle

Netizens can by means of the facilities provided by advanced interruptive technologies such as Big Data, Cloud Computing, IoT, AI and ML and so on conduct their daily activities more efficiently and effectively. These technologies enable integration of massive, scattered datasets, efficient interpretation of the integrated data, and speedier communication of the information. These facilities bring along obvious benefits and security threats on account of a huge amount of information being made available. As the cyber-world becomes a more transparent and open environment, netizens are better informed and able to innovate marketing, to accelerate business promotion, to enlarge data storage capacity and communication coverage, to increase retrieval facilities, and to improve transaction speed. At the same time, through that transparency loopholes in these new technologies where threats are bred provide opportunities for the cyber-miscreants to exploit and attack. However, as the clandestine activities that are brought to light, for example, the Snowden episode (*South China Morning Post*, 2013) and the Panama Papers leak (Wilson, 2016), the consequences can be beneficial to some people/organizations and adversary to others.

DOI: 10.4018/978-1-7998-3479-3.ch030

Figure 1. A conceptual graph of the circle and the phenomenon



The better-informed are the consumers, the more information is demanded and supplied. Supplying more information improves profit and uses more technologies. Using more technologies makes more information available. Consumers are happier and demand more information. More information is supplied and more technologies are used. The volume of information increases as more information is collected and generated. This culminates in a *vicious circle*.

Data Protection: A Chronic Problem

We spend more (the big spending on cybersecurity), we lose more (still get hacked and the enormous aftermath of the cyberwars) but data protection has to go on. The symptom is big spending and cyberwar and the cause is that we do not take ethics seriously.

The Big Spending: This the following exemplify: US\$1.5 billion by American financial institutions alone (US Homeland Security Research, 2016). £6.2m in 2016 compared to £3m in 2015 by UK organizations (CyLon, 2017). Gartner predicted in 2017 that worldwide spending to grow to \$93 billion in 2018 from \$86.4 billion in 2017 (Bradley, 2017), and forecasted in 2018 that Hong Kong's budget for data security (the fastest growing product category) would reach HK\$85 million in 2019, mainland China to reach almost 23.9 billion yuan (HK\$27.2 billion) and 27.3 billion yuan, respectively 2018 and 2019, and globally, expected spending on information security products and services to reach more than US\$114 billion in 2018 and to US\$124 billion in 2019 (Bushell-Embling, 2018).

Cyberwars symbolizes nation-to-nation conflicts such as the US-China trade-based and military-oriented cyberespionage (Onag, 2018), the alleged Russian interference of the US general election in 2016 (The New York Times, 2018a), and the UK allegation of Russia's tampering of Brexit voting (The New York Times, 2018b), also denotes conflicts at lower scales such as the fight between advertisers and users of ad blockers (Wicker and Karlsson, 2017) and the dispute between consumers of the digital technologies and

the cybermiscreants (Talevski, 2018). Cyber-ransom exemplifies another type of cyberwar, typified by the following three cases: WannaCry vs UK's National Health Service where dozens of computers were infected every minute and WannaMine vs the world, hijacking the computers around the world to mine a cryptocurrency called Monero (The Telegraph, 2017); and the case where the ransomware encrypted the files of infected users, and demanded for decryption US\$300 (HK\$2,300) in bitcoins within three days, US\$600 (HK\$4,700) after the three-day deadline, or all the files would be permanently encrypted after one week (in Hong Kong on 15 May 2017) (Hong Kong Free Press, 2017).

Why the futile result? Our first reaction would be to blame the data protection software. This is, however, not fair because the software is implemented according to the design so the fault should lie in the design, not the software. Then, the design itself should not be blamed because design is governed by the data protection policies (at corporate level) and security standards (at operation level). Therefore, the design which leads to unsatisfactory end-products is but a secondary cause at the next level. These policies and standards, by design, concentrate on ensuring a good return on investment of the technology (financial viability) and an optimal utilization of technologies acquired (technical efficiency) but are silent on the socio-ethical aspect of the problem. It follows that the software which is incapable of covering issues of a social/ethical nature is not to be blamed. The fault lies in the policies and standards. Does this mean that the primary cause is the policies and standards? Not quite; they are set by the decision-makers of the policies and standards who neglect the ethical elements in their deliberation and formulation. It follows that the ultimate cause is that we do not take ethics seriously.

Why don't we take ethics seriously? The problem is rooted in the way the data are collected about and are used adversely against data subjects and the *right* of the data subjects to that data, and has to deal the *techno-ethical-risk* which is originated in the way the technology is applied and the human users of the technology and the data, though the technology per se is neutral, yet the use of the technology is not (Parker, 1986; Neumann, 1995; and Williams, 1997). The protection of data privacy means to ensure that data privacy is not breached, that is, the data content is securely stored and accessed only with proper authority, and that the right of the data subjects not abused. Solving the problem means developing, reviewing, and implementing information protection and data privacy policies, standards, guidelines and processes (the policies), and mitigation measures; and to identify new technology risks; and to ensure that the policies are appropriately enhanced, communicated and complied with, and that mitigation measures are properly implemented. Also, our indifference to or lack of intrinsic interest in ethics, an attitude that most of us adopt, contributes to the primary cause. The muddled view of the basic concepts and the fallacious belief hinder a holistic understanding of and exacerbate our view. On many an occasion when ethics pops up in a conversation, many of us would say ethics is important but tend to obfuscate upon being pressed for elaboration. Not all of us, especially and more importantly the information technologists, have an intrinsic interest in ethics. The simple reason is our indifference to or lack of intrinsic interest in ethics, an attitude that most of us adopt. Muddled view and fallacious belief are the main culprits. A plain example is that on many an occasion when ethics pops up in a conversation among techies, the majority would say ethics is important but tend to obfuscate upon being pressed for elaboration. Not all of us, especially and more importantly the information technologists, have an intrinsic interest in ethics.

Security-Ethics-Law: A Tripartite Relationship

First, ethics and law are obviously closely related and mutually supportive of each other that the Donald Tsang case exemplifies. Law and ethics seem *separate* when the judge instructed the jury: "This is a legal

Figure 2. Ethics-Law-Security Chain (Source: Lee, 2015b)

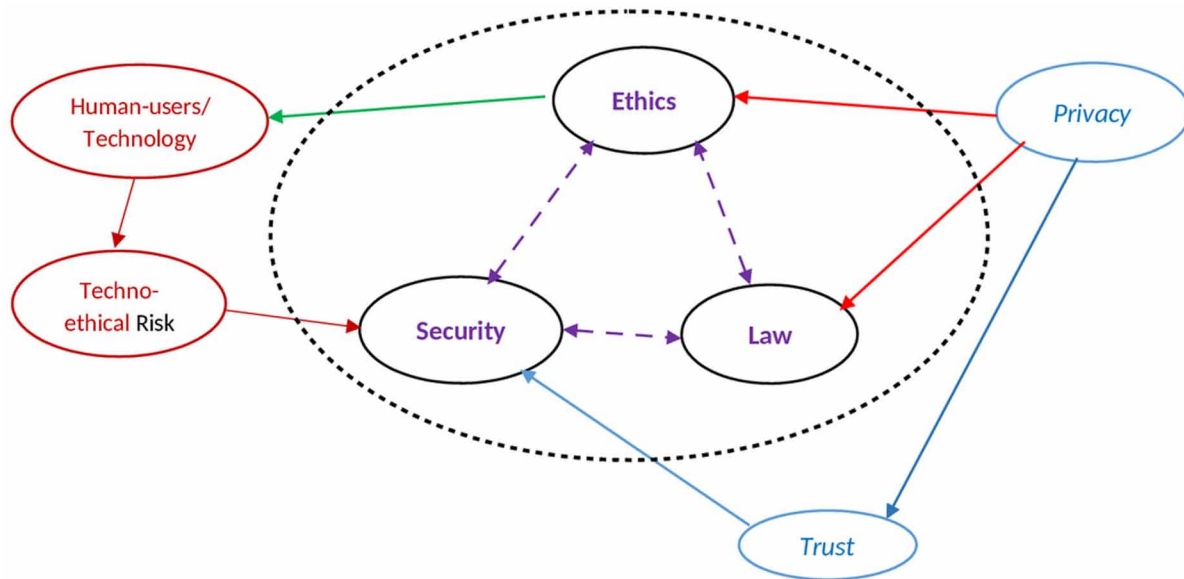
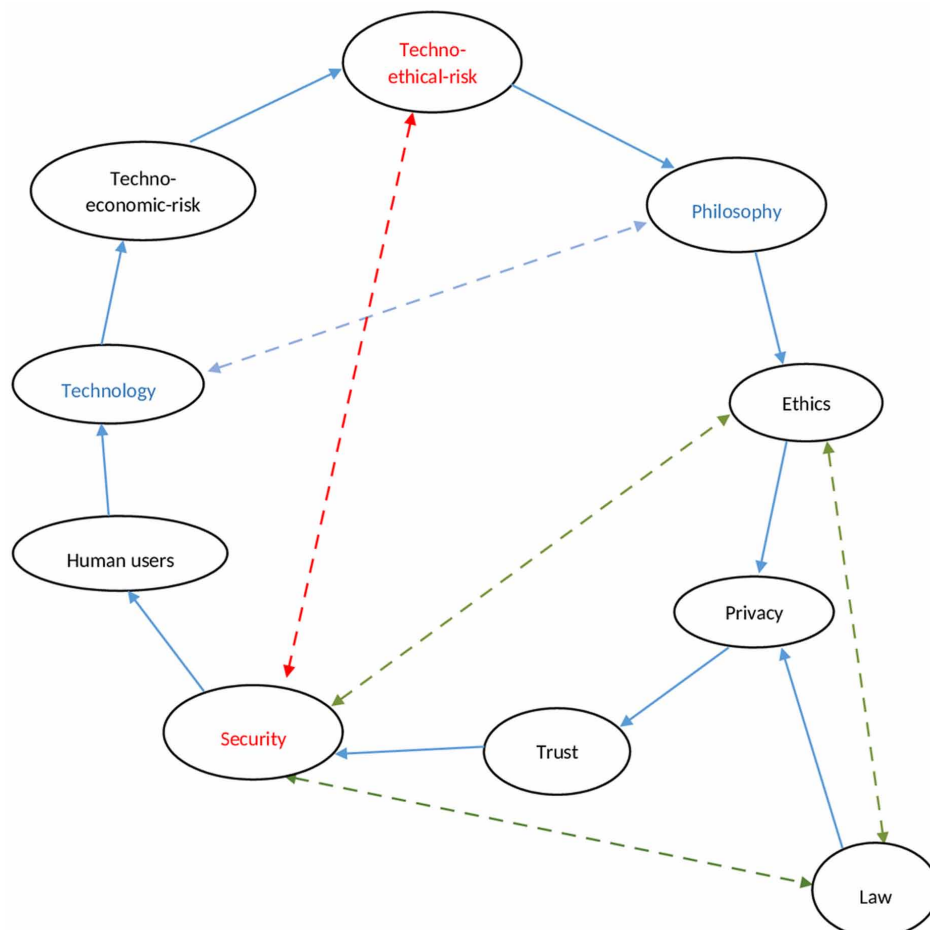


Figure 3. Connecting Philosophy/Ethics and Technology/Security/Risk (Source: Lee, 2015b)



case, not a moral case” yet law and ethics are *linked* when the defence counsel pleaded for a suspended sentence and a shorter term premised on two points: “lack of morally questionable motive” and “heavier sentence if motives morally questionable” (*South China Morning Post*, 2017; Lau and Hollingsworth, 2017). Second, ethics and security are interlinked. In cyberspace as in the physical world, security relies on trust; trust depends on privacy; and privacy is breached if law is not observed and if ethical principles are not followed. So, one won’t trust, or disclose, anything confidential to any who does not respect privacy; one would trust his or her security to the protecting agent who would act according to ethical principles. Hence, trust will be lost or damaged and repercussions will come after if there is no privacy. No trust, no security: no one will trust you if you act in contradiction to ethics or fail to obey the rule of law, and absence of trust will block privacy for no one will disclose information of a private nature to an untrustworthy recipient. (See Figures 2 and 3 which depict the Ethics-Law-Security Chain (Lee, 2015b): abuse of the technology (by human-users) will lead to risk (techno-ethical risk) as it violates ethical principles – certainly a security concern.

Risk, Ethics, Privacy: In a Muddle

Many of us habitually or subconsciously take risk for physical damage, financial loss or punishment by law; ethics for some intangible concern about right or wrong, or good or bad; and think of privacy as the same as about being let alone or free from interference. These notions are not wrong but not enough for a wholesome appreciation of dilemmas or aftermath of cybersecurity cases. Risk, ethics, privacy and concepts related to them are in some kind of muddle and must be cleared, else the status quo will deteriorate, not improve.

Risk is more than is more than physical damages, financial losses, and criminal sanctions. Taking such a view largely attributes to the orthodox decision model by Simon (1976), or due to misinterpretation of risk. Risk has been treated as a technical entity and measured in economic terms only when it is in fact a managerial concern and should be evaluated in social as well as technical terms, and technical, economic and social aspects of risk should be recognized and given equal priority and evaluated holistically. A shift of view is imperative since a risk is in fact a managerial concern – as ethical risk and audit risk exemplify – it should be evaluated in social as well as technical terms. Taking home from work a USB containing corporate confidential data (a company asset) for personal convenience runs the risk of breaking company rules and breaching data privacy or confidentiality. In this case, obviously the technology (the USB) is used in contradiction to ethical principles, and a risk is formed when that threat exploits the vulnerability where taking home an USB is a Vulnerability, taking home an un-encrypted USB containing confidential data is a Threat. It is a new type of risk created when called techno-ethical risk, or simply ethical risk. A risk can be speculative if the outcome of ethical actions will bring about gains and unethical actions will yield a loss, or dynamic if the chance of occurrence is associated with technological change or change in general including man-made change. For example, security threats regarding privacy and culture of Cloud Computing are speculative because of the opportunity for potential revenue due to additional protection charges for guarding against invasion of privacy and for provision of resolving cultural differences; they are obviously dynamic, not static, as they occur not through irregular action nature flood, earth-quake, and so forth.

Ethics is more than about “right or wrong else there won’t be complaint that ethics is vague or slippery and there won’t be so many various principles/theories. Fallacy plays a significant role: ‘What is legal is ethical’ is a fallacious. To distribute mailing lists without permission is legal but unethical for it is not explicitly prohibited by law but may step over the privacy of the people on the lists; to copy pro-

proprietary software as a backup, even when the agreement specifically prohibits copying for any purpose is ethical but illegal because it may be tolerable but may infringe the copyright law. ‘If you can do it, so can I’ is often heard but supposing ‘what you do’ is wrong, then I do wrong, too! This is a circular or relative argument that tell ‘what you do’ is right or wrong as ethical theories such as Relativism and the Golden Rule give no indication of an action being ‘right or wrong’. Hence, there is need for other theories such as those based on ‘duty’ and ‘results of action’. Further, ethics has an Invisible-Hand with a dual function with a double-edged sword effect. Diagnosing an action, ethics plays the role of an attacker by exposing unethical consequences if that action violates ethical principles and of a defender if the action is consistent with the ethical principles.

Privacy is not the same as being alone. For example, two people can have a private relationship that might be physical or intellectual. Stepping over the line of someone’s privacy is rude and ethically unacceptable but it affronts that person’s dignity, not that person’s privacy. However, it is a violation of someone’s privacy when that someone is treated as a means to an end. Privacy can, in simple terms, be viewed as a “zone of inaccessibility” or “right to be left alone”. The former defines the extent to which one can control who has access, meaning that one has privacy in this zone, and entering this zone without permission is a violation of privacy, and is prohibited; the latter may mean that privacy can be “being let alone” or “being free from intrusion” or “physically being let alone” or “being free from intrusion into one’s physical proximity”.

In a digital business context, data privacy or cyber-privacy relates closely to cybersecurity. Data privacy is a legal cum ethical issue that concentrates on data governance and data use; so it defines the organization’s rules and procedures for collecting, storing and using personally identifiable information (commonly abbreviated as PII) in a rightful way. Cyber-privacy (a particularly instance of cybersecurity) is a managerial, operational issue that focuses on security and demands certain actions to protect data; it takes actions according to those rules and procedures to protect data against unauthorized access, corruption and loss using encryption, secure communications protocols and measurable security policies. So, data privacy and data protection are certainly similar and often overlap but not always synonymous but not always equal, especially in the digital business context because it is extremely difficult to ensure privacy when data are not protected by technology. If someone can maliciously use an individual’s personal information, its privacy is wholly uncertain. It is important that organizations that act as the data processor and controller employ data protection technologies to fortify the governance processes of the data. To this end, many organizations consider taking data privacy audit (check and ensure data governance and classification, data security, and third-party contracts). Notwithstanding, data privacy should not only consider merely data protection but also contractual principles because when using Google, Baidu, Amazon, Alibaba, Facebook and others, we all agree with these companies’ terms and conditions. Data protection is a technical and economic thing (the cost of data privacy breaching for example) and requires technical actions to uphold accuracy, completeness, availability, etc of the data content, and to admit only authorized access, and so on. Contractual principle is a legal thing and must observe the laws of the land. This techno-legal definition tells us that data privacy is achieved when technical efficiency and economic/financial viability are met.

Given that ethics varies depending on culture, time and geography, so does privacy. This raises a string of questions about value. Privacy does not have a stable universal value, but it has an intrinsic, a core and a social value (which is regarded as commonly applicable universally to cases perceived to have these values); the techno-legal definition is deficient, thus data privacy should also consider social acceptability to cover cases where value depends local law in force, culture or habit in fashion at the time being, in addition to technical efficiency and legal viability.

RECOMMENDED SOLUTION

The proposed remedy is ethics-based, aiming at mitigation rather than eradication and comprising the following:

- Take ethics seriously
- Have a sound understanding of the applicable theories of ethics and be adept to linking actions to these theories
- Adapt to a shift of view of risk and ethics
- Carry out ethical analysis and be adept to applying the *Ethical Computing* tools that include *Ethical Matrix* and *Hexa-dimension Metric*

Ethical Computing Feasibility

The ethics-based remedy feasible on account of

- Keeping all hackers out is unrealistic or impossible.
- Any aftermath of cyberattacks that the like of Cambridge Analytica exemplifies cuts through all three domains of law, ethics and security.
- Ethical Computing, a practice grounded in Computer Ethics, is poised to supply the ethical doctrines (to persuade against wrong-doings) and to take ethical actions (to nurture trust) with the resultant effect to relax or soften the demand on data protection and gradually reduce the information security budget and the cost of cyberattacks
- The attitude that ‘*we don’t take ethics seriously*’ is automatically taken care of. The IT community generally recognizes that ethical issues exist, and considers them important. Reflective of this point and attesting recognition at national level of the importance of ethical issues is the establishment of the Data Ethics Council in June 2016 by the UK government “as a means of addressing the growing legal and ethical challenges associated with balancing privacy, anonymization, security and public benefit” due to the problem caused by Big Data. Another sign that ethical issues are getting the attention they deserve at international level is that EU enforced a more stringent data protection standard through its GDPR (in force from 25 May 2018).
- *The knowledge of ethics has the effect of guiding us to ethical use of the computer.* A seven-year exploratory study supports this hypothesis. The resulting statistics (Table 1) indicate that 8% of the students (including those matured age part-timers with over ten years professional experience) claimed that they were aware of Computer Ethics but not certain if they use the computer ethically at the start of the course, compared with 72% claimed that they understood ethics and would use the computer ethically at the end of the course, and the remaining 28% (post course population) claimed no change for various reasons. Overall, knowledge of ethics has a remarkable impact. It is noteworthy, however, ethics has no impact on 21% (post course population) who claimed that understanding ethics did not change how they use of the computer. The empirical data is consistent over the seven consecutive years ($p > .05$). (Lee & Chan, 2008)

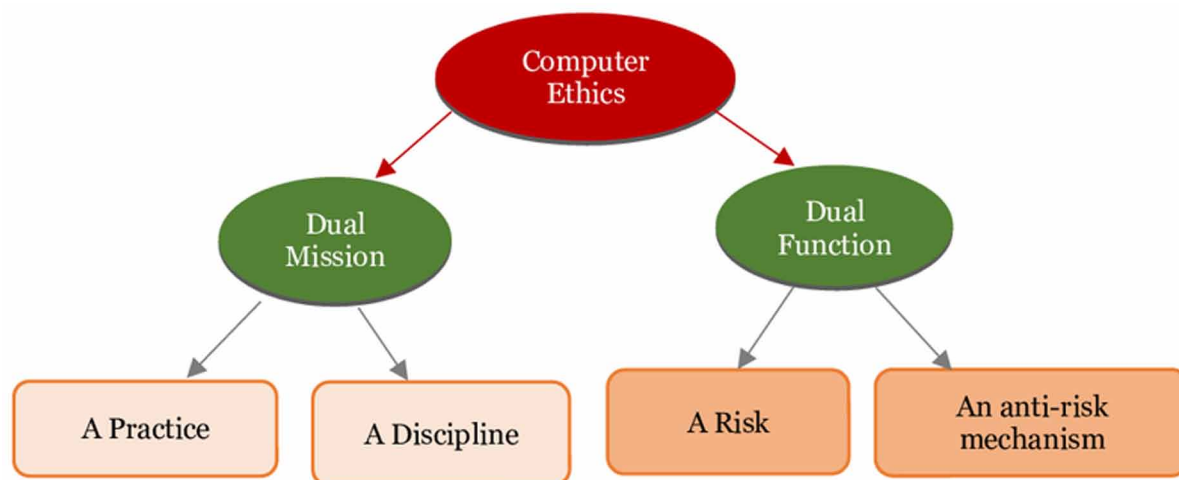
Ethical Computing

Ethical Computing is the practice of *Computer Ethics* that identifies/discovered ethical issues arising from developing and using computer-based application systems, and analyzes these issues to come up with balanced solutions. Its current major issues of interest include privacy, intellectual property, digital-divide, professionalism, trust, anonymity, cyberbullying and netiquette. Its methods and tools for ethical analysis include the *Ethical Matrix* and *Hexa-dimension Metric* among others (Lee, 2015a; Lee, 2019)

Computer Ethics

Ethics is same in peace times as during the *cyberwar era*. *Computer Ethics* or cyberethics, is a relatively new Applied Ethics. Defined slightly differently in the literature, notably Moor (1985) and Johnson (2009), the field is basically concerned with the *ethical issues* surrounding the computer. In this paper, Computer Ethics is viewed as a dual mission and a dual function (Figure 4) (Lee, 2014-15).

Figure 4. Conceptual Graph – Double Duality (Source: Lee, 2014-2015)



Ethical Matrix

This is a tool originally developed for the food and agriculture industry to carry out ethical analysis. The rows represent the stakeholders (relevant to the issue being studied); the columns correspond to the *prima facie* ethical principles respected for; and the cells contain the concerns (the main criterion that would be met if the specified ethical principle were respected for a particular stakeholder). The matrix is based on *prima facie* principles or rules of action that are valid at first sight. The advantage is that these principles allow a stronger case to overrule a weaker case in particular circumstances, thus facilitating ethical decision-making (because following the rules rigidly might lead to a deadlock or endorsing conflicting courses of action). (Mepham, etal, 2006, pp 8, 9)

Adapted for ethical analysis in the IT context, the columns of the matrix correspond to *values* held by stakeholders with respect to ethical principles, and the rows correspond to the stakeholders or inter-

est groups. The number of columns and rows varies as required. The cells contain the concerns of the stakeholders (the main criterion that should be met) with respect to the value held (by the stakeholders) or a particular ethical principle. (Lee, 2015a)

Hexa-Dimension Metric

This is developed as an aid to decision making in an IT context, aiming to supplement the variates that the orthodox decision model did not cater for and, as a checklist to ensure that the important aspects (six) are taken into consideration in the decision process. The six principles/measures are

- financial viability – to justify return on the investment on technology, that is, ensure that the actions taken are financially sound and viable consistent with the corporation's performance metric on operational goal, risk absorption standards and limits, etc. as advised by the Finance and Operations experts, and measured by ROI and NPV.
- technical effectiveness – to ensure of utilization the technologies is optimized, that is, use of up-to-date technology to meet targets effectively and efficiently
- legal validity – to ensure that the laws current in force and the organization's rules and conventions are abided at the advice of the corporate counsel or the Legal Department, and is measured by the extent and degree the law is obeyed
- ethical acceptance – to take actions that are consistent with ethical principles and the corporation's Public Relations and Human Resource Management policies using the Ethical Matrix method (Lee, 2015b) to assess the consistency with the relevant ethical principles of duty and reciprocity.
- social acceptability – to make sure that the benefit of all stakeholders including the public has been considered and deliberated consistent with local culture and habit and can be measured by survey results of the extent of public benefited based on the ethical principles of utility and consequence, and that the huge amount of data consumed and generated by the technology is safe and not abused
- ecological sustainability – to protect the environment since running the technologies consumes a lot of natural resources including electricity, water, paper, etc., and generates air, noise pollution and to take actions that are friendly (not harmful) to the environment aligned with the policies of energy/materials saving, noise and air pollution control, radiation emission rate and so on using, for example, EIA (environmental impact assessment) laid down by the Hong Kong Government

Incidentally, the GCF Computing (Patrignani & Whitehouse, 2014) correspondingly and independently cater for the six measures. Namely,

- **Good Computing** which targets at workable & technically efficient deliverables and applies the internet technologies, Big Data, Cloud Computing, Ubiquitous computing (Evasive Computing), etc, handles Legal, Technical & Financial efficiencies.
- **Clean Computing** which aims at energy & materials savings and applies Green Computing (environmentally responsible and eco-friendly use of computers and their resources – Technopedia), handles Environmental efficiency
- **Fair Computing** which is concerned with corporate social responsibility and applies Ethical Computing (Lee, 2015b), handles social and ethical efficiencies

Operationalizing

- Initialize the remedial procedure requires a 5-item precondition:
 - Board of Directors support with respect to corporate policy aspects
 - cooperation of the supporting infrastructures including the organization's human resources management, legal, finance, and ICT functional units with respect to technical support and reference
 - the inclusion of a mandatory appraisal of ethical consistency in conduct in an annual performance review
 - identify the relevant critical factors depending on the target end-users (corporate-wide or a functional unit, nature of operation – for example, environmental impact is critical for a mining company or a factory, but may probably be skipped for an Information Security Unit)
 - determine a scheme for quantifying the elements of each factor for measuring, prioritizing and balancing the factors. Depending on the attributes/factors determined, the steps to be taken to measure the effectiveness culminates in an operationalization scheme (Table 1)

Table 1. Schedule of operationalization

Generic				User specific	
Attributes/factors	Remarks	Control	Targets	Drivers (*)	Supported by
Legal validity	Abiding the law, rule & regulations currently in force	External	Local law and convention	Corporate counsel /legal dept.	Good Computing
Social desirability	Beneficial to all stakeholders including the general public	External	Local culture and habit; standards for perceivable benefits to primary stakeholders and the public		Fair Computing
Ecological sustainability	Friendly to natural and industrial/business environment	External	Policy/guidelines for energy/ materials saving; ecology related law (if any)		Clean Computing
Ethical acceptability	Consistent with ethical principles	Internal	Corporate social responsibility; corporate culture	PR, HRM	Fair Computing
Technical effectiveness	Use up-to-date technology to meet targets effectively and efficiently	Internal	Sufficient & necessary human & technological resources to support infrastructural support	ICT, HRM	Good Computing
Financial viability	financially sound and viable	Internal	Corporate policy (performance metrics: operational goal, risk absorption standards & limits, etc.)	Finance, Operations	Good Computing

ILLUSTRATION

The Octopus Case: Selling Customers' Personal Data

The Story

The Company sold cardholders' personal data and milked HK\$44 million+; the Privacy Commission investigated and prosecuted; some legislators demanded a legal investigation; and the CEO eventually resigned. (South China Morning Post. 2010). The CEO insisted that

- “She didn’t do anything wrong (nothing illegal)”.
- “Everybody is doing it”.
- “She had a duty to run a profitable operation”.
- “The money from sales has gone to the company profit, a positive return to shareholders”.

Ethical Analysis

So, she did nothing illegal. But is her action ethical? She made a circular argument relying on *Relativism* which is akin to saying “If Joe dumps his rubbish next door in Sue’s backyard, why can’t I do the same to Jenny, else I won’t be as smart as Joe if I don’t do that”.

So, she got a duty to perform, to make a profit for the company and the shareholders – a *deontological* (duties-based) argument. She had done her job well; the result was beneficial to the company and shareholders – a consequentialist (results-based) argument. The consequence may be good, but for whom, for how many people – herself only, the company and shareholders, the majority of the whole population of Hong Kong?

The proceeds from selling the data benefits the CEO and perhaps the company and shareholders, certainly not the cardholders, not to mention to rest of the population. This is certainly not consistent with the *utilitarian doctrine*. The cardholders have been cheated and treated as a means to the CEO’s want. This contradicts the *categorical imperative*. The CEO acted against the *Gold Rule* as she certainly won’t like to see her own personal data being collected by a third party without her knowledge and permission, and sold for a profit.

Lastly, collecting cardholders’ personal data inconsistent with the purpose and selling them without the data subjects’ consent defy the value cherished and cared by the majority of people – an action against *virtue ethics*.

Ethical Matrix Analysis

Table 2 displays the stakeholders identified, the ethical principles respected for, and the stakeholders’ concerns. The matrix reveals that

- The CEO’s main concern is obviously her job security, undoubtedly a duty for herself. Finding a way to improve or bring in a profit is natural and fine. However, that way must be consistent with the ethical principles. She fulfilled her duty; she did nothing illegally; she may or may have a clear conscience with respect to selling the cardholders’ personal data; she may not be looked at approvingly by the Golden Rule but she certainly deviated from the Categorical Imperative.

- Those cardholders who signed up for the Octopus Rewards were obviously attracted by the benefits that they were most interested in. This raises the issues responsibility and rights: Shouldn't they bear certain responsibility? Were they in a position to complain of being cheated?
- And more.

Table 2. Ethical Matrix - a First-cut Result

Values Stakeholders	Well-being (utilitarian/consequential)	Deontic Ethics	Virtue Ethics	Justice/Fairness	Golden rule	Categorical imperative
CEO	Job security	Performance as CEO	A clear conscience	Not stepping over the redline	Self-protection	Not appear to taking advantage of others
Cardholders	Enjoyment of benefits provided by the scheme	Materialize the benefits promised	Confidentiality of personal data	Fair, prima facie	Self-interest	Ensuring not to be exploited
The Company & shareholders	Profit & ROI	Performance monitoring	Company image and reputation	Meeting the set profit margin		
Regulators	Fair and just way trading	Proof of fair trading		Compliance of fair & proper trading		

Hexa-Dimension Metric

Table 3 shows that the problem does not involve technology or use of technology even though the use of the card is an application of information technology. Hence the technical and ecology measures are excluded from the analysis. However, as indicated, the ethical and social dimensions of the action (i.e., collecting and then selling personal data) must be addressed.

Further work to quantify the stakeholders' concern (expressed in the cells of the Matrix) and the verdicts, called indicators, remained undone and will be continued in future. The indicators will enable the users to prioritize the issues and the allocation of resource to rectify the deficient areas.

Table 3. Hexa-dimension Metric

The measures	Verdicts	Check
Financial viability	Making that extra income is financially viable for the CEO, the Company and shareholders but at the expense of the Octopus Rewards participants	Ö
Technical effectiveness	Not a technical issue; not applicable	n. a.
Legal validity	No prima facie evidence of illegal action	Ö
Ethical acceptability	Data collection and data use ethically questionable; unacceptable	×
Social desirability	Questionable in utilitarian view	×
Ecological sustainability	Not an ecology issue; air pollution or climate are non-issue; obviously not applicable	n. a.

FUTURE RESEARCH DIRECTIONS

The concept that a decision should meet the six criteria is new, but each criterion in the hexa-dimension metric model is not new. The metric has the potential to serve others purposes, for example, Management Science: the metric can be used for enhancing identification and evaluation of the relevant factors of the traditional and popularly adopted decision model like the rationality and logic based Simon model (1976). Such decision models are no longer sufficient because decision makers are not always rational and sometimes non-rational, or judgmental decision making based procedural or bounded rationality (Gigerenzer, 2001) and escalation of commitment decision making that “under certain conditions, decision makers who make an initial decision become overly committed to the original choice and then subsequently make decisions biased by psychological commitment” (McCarthy, Schoorman and Cooper (1993, p. 9)). Also, that the decision variables are formulated and measured mainly in economic terms must be updated as they are also legal, ethical, social, and more recently, ecological in nature. This raises several questions for future work:

- Quantify the results provided by the Matrix and the Metric – developing a method for evaluating
- Field-test the efficiency of Hexa-dimension Metric for designing a code of practice in an IT organization using a prototype. On-site study of its effectiveness as a handy reference at the different levels of the corporation (Lee, 2019)
- An empirical study of the Metric to enhance the chance and scope of finding factors likely to bring about ethical, social and ecological repercussion on decision making

CONCLUSION

The information security professionals are in urgent need of effective and pragmatic guidance as the information security function in a technology-driven information-intensive environment becomes more complicated and harder due to the new socio-techno risks. Data privacy protection becomes a primary concern as privacy breach occurs more frequently and leads to devastating consequences to the individuals, corporations and governments. The newly developed framework, though at this embryonic stage of growth still awaiting empirical evidence and popular acceptance in the IT arena, poises to play a useful role, to release the burden on as the extant policies and standards including the privacy laws such as UK’s DPA (UK Government, 2018), EU’s GDPR (GDPR, 2019), and Hong Kong’s PDPO (Office of the Privacy Commissioner for Personal Data, Hong Kong, 2019).

To operationalize the ethics-based mitigation requires the Board’s mandate and inter-departmental support, identification of the relevant critical factors depending on the target end-users, a scheme for quantifying the elements of each factor for measuring, prioritizing and balancing the factors. As illustrated, Ethical Matrix is useful in providing a holistic view of the stakeholders and the issues of concern with respect to principles they value, and Hexa-dimension Metric in checking an action against the six measures, hopefully to reach a more complete and balanced conclusion.

It is worthy of note that the Hexa-dimension Metric is an initiative prompted by the phenomenon that no matters how meticulous is the decision-making process, the results are seldom satisfactory. Why such a phenomenon appears at all is akin to “man doesn’t live on bread alone” or decisions that are argued in rational, logical and financial terms only, and is reflective of the Simon Model, the guiding light for decision making which is deep-rooted in our thought and practice of management, being congenitally

defective. The fault lies in some kind of self-contradiction for the Simon doctrine assumes that decision making is rational and objective and the decision variables are considered in financial terms only yet the fact is that decision makers are not always rational and sometimes judgmental or reliant on escalation of commitment, or emotional, and risk and cost can be ethical, social, legal, technical and ecological in nature. A shift of view of risk and cost is advisable.

Referring to the chronic problem where hefty investment in the expensive technologies, it is a matter of course to expect a justifiable return, and an optimized rate of utilization of the expensive technologies. To measure success at this point in financial and technical terms is essential. Since the huge amount of information generated can be abused, both legal and ethical issues arise. The use of the information may serve well some, but not some other, stakeholders, so immediately emerged are social issues. As more technologies are used, thus more natural resources consumed, impact on the environment must be considered. In sum, to arrive at a satisfactory solution, decision makers must consider financial viability, technical effectiveness, legal validity, ethical acceptance, social desirability and ecological sustainability. To balance or strike a trade-off among the six attributes and measure the outcome thereof, the Hexa-dimension Metric is recommended. Hence, the Hexa-dimension Metric is a good instrument not just for privacy protection policy formulation but also for clarifying/determining pragmatically “ethical” and “effective” in ethical leadership and effective leadership which all professionals would strive for.

REFERENCES

- Bushell-Embling, D. (2018). *HK information security spend to reach HK\$3.5b this year*. <https://www.cw.com.hk/it-hk/hk-information-security-spend-to-reach-hk-3-5b-year>
- CyLon. (2017). *Increased corporate spending on cybersecurity, but to what avail?* <https://blog.cylonlab.com/increased-corporate-spending-on-cybersecurity-but-to-what-avail-1f8f46934b34>
- GDPR. (2019). *General Data Protection Regulation*. <https://gdpr-info.eu/art-5-gdpr/>
- Gigerenzer, G. (2001). *Decision Making: Nonrational Theories*. In *International Encyclopedia of the Social and Behavioral Sciences* (Vol. 5). Elsevier.
- Hong Kong Free Press. (2017). *3 cases of WannaCry ransomware virus reported in Hong Kong so far, says official*. <https://www.hongkongfp.com/2017/05/15/3-cases-wannacry-ransomware-virus-reported-hong-kong-far-says-official/>
- Hong Kong Government. (n.d.). *A guide to the EIA (Environmental Impact Assessment) Ordinance*. Retrieved on April 25, 2016, from <http://www.epd.gov.hk/eia/english/guid/index1.html>
- Lau, C., & Hollingsworth, J. (2017). *Former Hong Kong leader Donald Tsang taken to hospital while in custody*. <https://www.scmp.com/news/hong-kong/law-crime/article/2072324/prosecutors-push-retrial-donald-tsang-case-after-jury>
- Lee, W. W. (2015a). Ignorant of ethical pitfalls can hurt information security: Implications for IT users & Information Security Professionals. Special Seminar on Information Security, the Hong Kong Jockey Club.
- Lee, W. W. (2015b). Ethical Computing. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (3rd ed., pp. 2991–2999). doi:10.4018/978-1-4666-5888-2.ch292

Lee, W.W. (2014-15). *Ethical, Legal & Social Issues*. Lecture Notes, Postgraduate Diploma in eHealth Informatics, The University of Hong Kong.

Lee, W. W. (2019). Hexa-Dimension Code of Practice for Data Privacy Protection. *Advanced Methodologies and Technologies in System Security, Information Privacy, and Forensics*. <https://www.igi-global.com/chapter/ethical-computing-continues-from-problem-to-solution/213652>

Lee, W. W., & Chan, K. C. C. (2008). Computer Ethics: a Potent Weapon for Information Security Management. *Information Systems Audit & Control Journal*. www.isaca.org/jonline

McCarthy, A. M., Schoorman, F. D., & Cooper, A. C. (1993). Reinvestment decision by entrepreneurs: Rational Decision-making or escalation of commitment? *Journal of Business Venturing*, 8(1), 9–24. doi:10.1016/0883-9026(93)90008-S

Moor, J.H. (1985). What is Computer Ethics? *Metaphilosophy*, 16(4), 266–275. doi:10.1111/j.1467-9973.1985.tb00173.x

Moore, P. G., Thomas, H., Bunn, D. W., & Hampton, J. M. (1976). *Case Studies in Decision Analysis*. Marmonsworth. Penguin Books.

Neumann, P. G. (1995). *Computer Related Risk*. ACM Press /Addison-Wesley.

Office of the Privacy Commissioner for Personal Data. Hong Kong. (n.d.). *The Personal Data (Privacy) Ordinance, Chapter 486*. <https://www.pcpd.org.hk/english/files/pdpo.pdf>

Onag, G. (2018). *China's cyber spies hunt for business information*. <https://www.cw.com.hk/security/china-s-cyber-spies-hunt-for-business-information>

Parker, D. B. (1986). Computer Crime. *Financial Executive*, 2(12), 31–33.

Patrignani N. & Whitehouse, D. (2014). Slow Tech: a quest for good, clean and fair ICT. *Journal of Information, Communication and Ethics in Society*, 78-9.

Simon, H. A. (1976). *Administrative Behaviour* (3rd ed.). The Free Press.

South China Morning Post. (2010). <https://www.scmp.com/article/720620/octopus-sold-personal-data-customers-hk44m>

South China Morning Post. (2013). <http://www.scmp.com/news/hong-kong/article/1259422/edward-snowden-let-hong-kong-people-decide-my-fate>

South China Morning Post. (2017). <http://www.scmp.com/news/hong-kong/law-crime/article/2072324/prosecutors-push-retrial-donald-tsang-case-after-jury>

Talevski, J. (2018). *VMware's Pat Gelsinger: Security is broken*. <https://www.cw.com.hk/security/vmware-s-pat-gelsinger-security-broken>

Technopedia. (n.d.). <https://www.techopedia.com/definition/14753/green-computing>

The Telegraph. (2017). *NHS cyberattack: Everything you need to know about 'biggest ransomware' offensive in history*. <https://www.telegraph.co.uk/news/2017/05/13/nhs-cyber-attack-everything-need-know-biggest-ransomware-offensive>

UK Government. (2018). <https://www.gov.uk/government/collections/data-protection-act-2018>

Wicker, S. B., & Karlsson, K. (2017). Internet Advertising: Technology, Ethics, and a Serious Difference of Opinion. *Communications of the ACM*, 60(10), 70–77. doi:10.1145/3048384

Williams, K. (1997). Safeguarding Companies from Computer/Software Fraud. *Management Accounting*, 78(8), 18.

Wilson, S. (2016). The Panama Papers scandal. *MoneyWeek*. <http://moneyweek.com/the-panama-papers-scandal>

ADDITIONAL READING

Lee, W. W. (2014). Why Computer Ethics Matters to Computer Auditing. *ISACA Journal*, 2, 48–52.

Moore, P. G., & Thomas, H. (1976). *The Anatomy of Decisions*. Marmonsworth. Penguin Books.

Pazowski, P. (2015). *Green computing: latest practices and technologies for ICT sustainability*. Retrieved February 13, 2016, from www.toknowpress.net/ISBN/978-961-6914-13-0/papers/ML15-377.pdf

KEY TERMS AND DEFINITIONS

Chronic Problem: Refers to the situation that we spend more (the big spending on cybersecurity), we lose more (still get hacked and the enormous aftermath of the cyberwars) but data protection has to go on.

Ethical Analysis: Is about considering the competing factors when making ethical decisions because, at any one time, our behaviour may be affected by these factors which may be physiological, legal and spiritual. In common practice are three approaches: legal, formal or informal guidelines, and four principles: harm minimization, deontological (rights and duties) and consequentialist (outcome of the action) principles and Kantian categorical imperative (consistency and respect). Tools and methodologies in use typically linear menu-driven list consisting of four or five steps. Ethical Matrix was introduced but not yet popular until lately (see the next item). In any case, the aim of analysis is to identify the stakeholders and the concerns arising as a consequence of an action with respect to some ethical principles that the stakeholders value.

Ethical Matrix: Is a tool for ethical analysis adapted from Food and Agriculture to the IT context. The columns of the matrix correspond to *values* held by stakeholders with respect to ethical principles, and the rows correspond to the stakeholders or interest groups. The number of columns and rows varies as required. The cells contain the concerns of the stakeholders (the main criterion that should be met) with respect to the value held (by the stakeholders) or a particular ethical principle.

Ethics: Is simply put about right and wrong. But it is more than that. Deontic ethics is about duty-based actions taken according to principles of duty and rights; utilitarian ethics is on result-based action taken consistent with the consequence and utility doctrines. Judging an action being ethical may refer the gold rule, categorical imperative, virtue ethics, etc.

Ethics-Based Remedy: Aims at mitigation rather than eradication and comprises the following: take ethics seriously, have a sound understanding of the applicable theories of ethics and be adept to linking actions to these theories, adapt to a shift of view of risk and ethics, carry out ethical analysis and be adept to applying the ethical computing tools that include ethical matrix and hexa-dimension metric.

Hexa-Dimension Metric: Is a checklist to ensure that the important aspects (six) are taken into consideration in the decision process. The six principles/measures are financial viability, technical effectiveness, legal validity, ethical acceptance, social acceptability, and ecological sustainability.

Privacy: Is not the same as being alone; it is about one's "zone of inaccessibility" or "right to be left alone". Intruding into that zone is a violation of privacy and ignoring that right is an act of bullying, both are subject criminal charges.

Risk: Is more than is more than physical damages, financial losses, and criminal sanctions. Using the technologies in contradiction to ethical principles is a techno-ethical risk or ethical risk. There are also speculative risk (if the outcome of ethical actions will bring about gains and unethical actions will yield a loss), dynamic risk (if the chance of occurrence is associated with technological change or change in general including man-made change).

Tripartite Link: Symbolizes the interrelationship among Ethics, Law and Security where security relies on trust; trust depends on privacy; and privacy is breached if law is not observed and if ethical principles are not followed.

Vicious Circle: Depicts the phenomenon: The better-informed are the consumers, the more information is demanded and supplied. Supplying more information improves profit and uses more technologies. Using more technologies makes more information available. Consumers are happier and demand more information. More information is supplied and more technologies are used. The volume of information increases as more information is collected and generated.

A Utility Theory of Privacy and Information Sharing

Julia Puaschunder

The New School, USA

INTRODUCTION

Utility theory is concerned with people's choices and decisions based on preferences and values (Fishburn, 1968). Representing satisfaction experienced, utility is derived from the self-attributed worth and goodness of an option compared to other options. Standard neo-classical economic theory describes utility as a set of internally-consistent assumptions about options in the wish to maximize utility (Fishburn, 1968). Utility theory has leveraged as one of the most dominant theories in economics as an underpinning of rational choice and game theory. Utility is usually revealed in people's willingness to pay different money amounts for different options, leading to the concept of revealed preferences (Samuelson, 1937).

Whereas economic utility studies primarily focus on prescriptive approaches to guide how people should behave to maximize their well-being (Arrow, 1951, 1958; Majumdar, 1958; Simon, 1959); decision sciences started capturing how people actually decide regarding choices in an uncertain world and over time (Becker & McClintock, 1967; Edwards, 1954, 1961; Luce & Suppes, 1965). Expected utility theory introduces a first temporal discussion of expectations of utility rather than the actual utility derived from a choice (Alchian, 1953; Marschak, 1950; Strotz, 1953). Von Neumann and Morgenstern (1953) introduced that outcomes of choices are not known with certainty but have probabilities of occurrence, which weighted linear combination allows inferences about the overall utility derived over time.

Since the end of the 1970ies, a wide range of psychological, economic and sociological laboratory and field experiments proved human beings deviating from rational choices as standard neo-classical profit maximization axioms to fail to explain how human actually behave (Kahneman & Thaler, 1991). Human beings were shown to use heuristics in the day-to-day decision making as mental short cuts that enable to cope with information overload in a complex world (Bazerman & Tenbrunsel, 2011; Kahneman & Tversky, 1979; Thaler & Sunstein, 2008).

As one of the most recent developments in utility theory studies, behavioral economics find human utility choices biased (Bowles, 2004; Camerer, Loewenstein & Rabin, 2004; Ebert & Prelec, 2007; Kahneman, 2011; Okada & Hoch, 2004; Putnam, 2002; Sen, 1971, 1993, 1995, 1997, 2002a; Zauberman, Kim, Malkoc & Bettman, 2009) by heuristics (Kahneman, Slovic & Tversky, 1982; Simon, 1979), analogical thinking (Colinsky, 1996; Gentner, 2002), and minimized effort (Allport, 1979; Shah & O'penheimer, 2008).

In particular, people's cognitive capacities to consider future outcomes in today's decisions are limited (Doyle, 2013; Laibson, 1997; Loewenstein, 1992; Milkman, Rogers & Bazerman, 2009; Read, Loewenstein & Kalyanaraman, 1999; Read & van Leeuwen, 1998). Laibson's (1997) hyperbolically decreasing discounting functions more accurately describe choice behaviors of individuals, who tend to be impatient for smaller rewards now rather than waiting for larger ones later (e.g., Ainslie, 1992; Becker & Murphy, 1988; Doyle, 2013; Estle, Green, Myerson & Holt, 2007; Frederick, Loewenstein & O'Donoghue, 2002; Green, Fry & Myerson, 1994; Green & Myerson, 2004; Hansen, 2006; Henderson & Bateman, 1995;

DOI: 10.4018/978-1-7998-3479-3.ch031

Kirby, 1997; Kirby & Marakovic, 1995; Laibson, 1997; Loewenstein & Prelec, 1993; Mazur, 1987; Meyer, 2013; Murphy, Vuchinich & Simpson, 2001; Myerson & Green, 1995; Rachlin, Raineri & Cross, 1991; Sterner, 1994). Dynamically inconsistent preferences reverse as people are patient when deciding for the future and impatient when choosing for now (Hornsby, 2007; Laibson, 1997; McClure, Ericson, Laibson, Loewenstein & Cohen, 2007; Meyer, 2013; Reed & Martens, 2011; Thaler, 1981).

Field and laboratory experiments provide widespread empirical evidence for hyperbolic discounting and self-control failures (Frederick et al., 2002; Hoch & Loewenstein, 1991; Sen, 1971, 2002b) on money management (Alberini & Chiabai, 2007; Chabris, Laibson & Schuldt, 2008; Collier & Williams, 1999; Harrison, Lau & Williams, 2002; Keller & Strazzer, 2002; Kirby & Marakovic, 1995; Laibson, 1997; Laibson, Repetto & Tobacman, 2003; Salanié & Treich, 2005; Slonim, Carlson & Bettinger, 2007; Thaler & Shefrin, 1981; Warner & Pleeter, 2007), financial benefits (Cairns & van der Pol, 2008), credit card debt (Meier & Sprenger, 2010; Shui & Ausubel, 2004), medical adherence (Trope & Fishbach, 2000), public health (Bosworth, Cameron & DeShazo, 2006; Cameron & Gerdes, 2003; Chapman, 1996; Duflo, Banerjee, Glennerster & Kothari, 2010; Horowitz & Carson, 1990; van der Pol & Cairns, 2001), addiction (Badger, Bickel, Giordano, Jacobs, Loewenstein & Marsch, 2007; Becker & Murphy, 1988; Heyman, 1996; Laux & Peck, 2007; Madden, Bickel & Jacobs, 1999; Petry & Casarella, 1999), social security (Mastrobuoni & Weinberg, 2009), fiscal policies (Keeler & Cretin, 1983), commitment (Duflo, Kremer & Robinson, 2008; Sen, 1977, 2002b), health exercise (DellaVigna & Malmendier, 2004, 2006), employment (DellaVigna & Paserman, 2005), procrastination (Reuben, Sapienza & Zingales, 2010), diet (Read & van Leeuwen, 1998), subscription discipline (Oster & Scott-Morton, 2005), animal care (Green et al., 1994; Mazur, 1987), and consumption (Milkman, Rogers & Bazerman, 2008; Read et al., 1999; Wertenbroch, 1998). Failures to disciplinedly stick to plans for giving in to immediate desires (Ainslie & Haslam, 1992; Read, Frederick & Airoldi, 2012; Strotz, 1956) are explained by people caring less about future outcomes in the eye of future uncertainty (Luce & Raiffa, 1957; Shackle, 1955), perceived risk (Mas-Colell, Whinston & Green, 1995), and transaction costs (Chung & Herrnstein, 1967; Epper, Fehr-Duda & Bruhin, 2011; Frederick et al., 2002; Kirby & Herrnstein, 1995; Mazur, 1987; Read, 2001). Presenting temporal snapshots for now and later concurrently helps overcome myopia and decision-making fallibility (Puauschunder & Schwarz, 2012). In all these studies missing is an investigation of human decision making on preferences for information sharing and privacy.

Although communication and non-communication are day-to-day decisions of individuals; to this day, there is no stringently tested utility theory of information sharing and privacy. We lack a coherent decision science framework about when people choose to share information and when they rather want to stay silent for the sake of privacy. From the economic perspective, information sharing may impose temporal irreversible lock-ins or tipping points. The point of information sharing may be a reference point, in which one bit of more communication gives less utility than one bit of less information shared, hence one bit of more privacy, grants more utility in the sense of Kahneman & Tversky's (1979) behavioral decision science finding 'losses loom larger than gains.' There may also be a marginal decreasing utility derived from one bit more information shared but an exponential marginal utility gain from one more unit of information received given the fact that information can be put into context and an exponentially increasing marginal utility of information. Education, for instance, is the only good with an exponential marginal utility increase, as the more information one holds, the more complex connections one can make and use.

In the past, communication was depicted to decentralize organizations (Crémer, Garicano & Prat, 2007). Media was initially promoted to offer means of information transfer, political participation and protection against political abuse (Delli Carpini & Keeter, 1989; Neuman, Just & Crigler, 1992; Norris

& Sanders, 2003; Prat & Strömberg, 2005; Snyder & Strömberg, 2010). Evidence suggests that media coverage increases voter information, which increases the responsiveness of votes to policy, which increases the effort and selection of politicians, thus producing better policies (Prat & Strömberg, 2013). Media thus traditionally was portrayed as helping to keep politicians accountable (Prat & Strömberg, 2013). Media coverage was found to improve selection and incentives of politicians alongside voting responsiveness (Iyengar & Kinder, 1987; Snyder & Strömberg, 2010).

Critical studies in this regard show that there are negative downsides of transparency (Prat, 2005). Mass media can also erode social capital, as they potentially isolate people from real-world experiences (Olken, 2009; Putnam, 2000). A positive relation between federal funds per capita allocation to areas where the media covered political parties in power was found (Snyder & Strömberg, 2010). Research has been done on the effect of conventional media on politics including a nomenclature of biases that impose problems – especially against minority opinions (Prat & Strömberg, 2013). Ideological biases are found in conventional media and media effects captured on vote choice (Prat & Strömberg, 2013). While the negative facets of information on elections and the role of social media on voting outcomes has been widely discussed recently; yet to this day no stringent theoretical or empirical framework for the utility of privacy and information sharing on social media exists.

In the digital age, to study the trade-off between information sharing and privacy has leveraged into unprecedented importance. Social media revolutionized human communication around the globe. As never before in the history of humankind, information about individuals can be stored and put in context over time and logically placed within society thanks to unprecedented data conservation and computational powers. The big data era, however, also opened gates to unprecedentedly reap benefits from information sharing and big data generation (Puaschunder, 2017). The so-called nudigital society was recently introduced, shedding light onto the undescribed hidden social class division between social media users and social media providers, who can benefit from the information shared by social media users. Social media users share private information in their wish to interact with friends and communicate to public. The social media big data holder can then reap surplus value from the information shared by selling it to marketers, who can draw inferences about consumer choices. The big data can also be used for governance control purposes, for instance border protection and tax compliance control.

Drawing from the economic foundations of utility theory, this paper seeks to introduce the first application of utility theory to a preference-values predicament between communication and privacy in the new media era. Behavioral economics insights are advanced in shedding novel light on the conflict between the humane wish to communicate now versus combined information held by unknown big data compilers in the future. An exponential loss of privacy and hyper-hyperbolic risks in the future for the information sharer are introduced as behavioral economic decision-making fallibilities. For the overconfident information sharer, it remains largely unforeseeable what the sum of the individual information sharing tranches can lead to over time and what information its Gestalt holds for those who have big data insights over time, which can also be analyzed in relation to the general population. Governance gains a critical stance on new media use for guiding on public concerns regarding privacy and information sharing in the digital age (Puaschunder, 2017). While there is some literature on the history of media on politics (Prat & Strömberg, 2013), the wide societal implications of fake news and discounting misinformation has widely been overlooked in contemporary behavioral economics research and the externalities literature. Social sciences literature on privacy and information sharing has to be reconsidered in the age of social media.

The article is structured as follows: An introduction of the theory of utility and communication and information sharing is followed by an outline of the impetus of the digital big data age on privacy. The

first utility theory of information sharing and privacy will be theoretically introduced. Hyperbolic decision making fallibility will become the basis of argumentations around hyper-hyperbolic discounting – the novel argument that information sharing in tranches may lead to an underestimation of the privacy infringements when these bits of information can be put together over time and are compared to big data in order to infer about the individual in relation to the general population.

The subjective additive utility of information-shared tranche by tranche may underestimate the big data holder's advantage to reap benefits from information shared. The discussion introduces problems of the contemporary nudigital society (Puaschunder, 2017), in which big data compilers can reap a surplus value from selling compiled information (The New York Times, November 14, 2017) or manipulate vulnerable population segments based on their previously shared information (The Economist, November 4, 2017). Implications lead to open questions about ethics in the information age and recommendations for a reclaiming of the common good of shared knowledge in education about information sharing in the digital age as well as the democratization of information. Challenging contemporary behavioral insights theory aims at fostering a more informed, self-determined and protected digital society in the wish to uphold e-ethics in the 21st century big data social media era.

BACKGROUND

The wish for communication is inherent in human beings as a distinct feature of humanity. Leaving a written legacy that can inform many generations to come is a humane-unique advancement of society. At the same time, however, privacy is a core human value. People choose what information to share with whom and like to protect some parts of their selves in secrecy. Protecting people's privacy is a codified virtue around the world to uphold the individual's dignity. Yet to this day, no utility theory exists to describe the conflict arising from the individual preference to communicate and the value of privacy.

The Humane Preference for Communication

The act of conveying intended meanings from one entity or group to another through the use of mutually understood signs and semiotic rules is the act of communication. Communication is a key feature of humans, animals and even plants (Witzany, 2012). Steps inherent to all human communication are the formation of communicative motivation and reason, message composition as further internal or technical elaboration on what exactly to express, message encoding, transmission of the encoded message as a sequence of signals using a specific channel or medium, noise sources influencing the quality of signals propagating from the sender to one or more receivers, reception of signals and reassembling of the encoded message from a sequence of received signals, decoding of the reassembled encoded message and interpretation or sense making of the presumed original message (Shannon, 1948). Information sharing implying giving up privacy is at the core of communication. Communication can be verbal and non-verbal. Comprising very many different domains ranging from business, politics, interpersonal, social to mass media; communication is a humane-imbued wish and center core of every functioning society.

In society, language is used to exchange ideas and embody theories of reality. Language is the driver of social progress (Orwell, 1949). Linguists find discourse and information sharing inseparable from socio-economic societal advancement (Fowler, Hodge, Kress & Trew, 1979). Language and communication modes are implicit determinations of social strata (Orwell, 1949). Different institutions and media

sources have different varieties of language and information sharing styles. Access to information is related to social status and market power. Social visibility is a powerful and cheap incentive to make people contribute more to public goods and charities and be less likely to lie, cheat, pollute or be insensitive and antisocial (Ali & Benabou, 2016). Information receipt is an implicit determinant to classify and rank people to assert institutional or personal status in society (Fowler et al., 1979). Mass communication echoes in economic cycles in the creation of booms and busts (Puaschunder, work in progress). Media is also a hallmark of propaganda and political control (Besley & Prat, 2006; Prat & Strömberg, 2013). At the same time, privacy is a human virtue around the world.

Privacy as a Human Virtue

Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby share information about themselves selectively. The right to privacy grants the ability to choose which information about parts of the self can be accessed by others and to control the extent, manner and timing of the use of those parts we choose to disclose. Privacy comprises of the right to be let alone, the option to limit the access others have to one's personal information and secrecy as the option to conceal any information about oneself (Solove, 2008).

The degree of privacy varies in autonomy levels throughout individualistic and collectivism cultures. While the boundaries and contents protected and what is considered as private differ widely among cultures and individuals, the common sense in the world is that some parts of the self should be protected as private.

Privacy has a valued feature of being something inherently special or sensitive to a person, which can create value and specialty if shared with only a selected person or group. The domain of privacy partially overlaps with security, confidentiality and secrecy, which are codified and legally protected throughout the world, mainly in privacy laws but also in natural laws of virtues of integrity and dignity. Privacy is seen as a collective core human value and fundamental human right, which is upheld in constitutions around the world¹ (Johnson, 2009; Warren & Brandeis, 1890).

In personal relations, privacy can be voluntarily sacrificed, normally in exchange for reciprocity and perceived benefits. Sharing private information can breed trust and bestow meaningfulness to social relations. Giving up privacy holds risks of uncertainty and losses, which are undescribed in economics and in particular the behavioral economics literature on intertemporal decision-making (Gaudeul & Giannetti, 2017). People tend to be more willing to voluntarily sacrifice privacy if the data gatherer is seen to be transparent as to what information is gathered and how the information will be used (Oulasvirta, Suomalainen, Hamari, Lampinen & Karvonen, 2014). Privacy as a prerequisite for the development of a sense of self-identity is a core of humanness (Altman, 1975). Privacy is often protected to avoid discrimination, manipulation, exploitation, embarrassment and risks of reputational losses, for instance, in the domains of body parts, home and property, general information of private financial situations, medical records, political affiliation, religious denomination, thoughts, feelings and identity.

Technological shocks have a history of challenging privacy standards (Warren & Brandeis, 1890). The age of instant messaging and big data, however, has leveraged the idea of privacy to another dimension. The concept of information privacy has become more significant as more systems controlling big data appear in the digital age. With advances in big data, face recognition, automated license plate readers and other tracking technologies, the upholding privacy and anonymity has become increasingly expensive and the cost is more opaque than ever before (Ali & Benabou, 2016).

Privacy in The Digital Big Data Era

The amount of big data stored each second has reached an all time high in the digital era. Internet privacy is the ability to determine what information one reveals or withholds about oneself over the internet, who has access to personal information and for what purpose one's information may be used. Privacy laws in many countries have started to adapt to changes in technology in order to cope with unprecedented constant information surveillance possibilities, big data storage opportunities and computational power peaks. For instance, Microsoft reports that 75 percent of U.S. recruiters and human-resource professionals use online data about candidates, often using information provided by search engines, social-network sites, photo and video sharing tools, personal web appearances like websites and blogs, as well as Twitter.

Social media tools have become large-scale factories with unpaid labor (Puaschunder, 2017). For instance, Facebook accounts for the largest social-network site with nearly 1,490 million members, who upload over 4.75 billion pieces of content about their lives and that of others daily. The accuracy of this information also appears questionable, with about 83.09 million accounts assumed to be fake. Aside from directly observable information, social media sites can also easily track browsing logs and patterns, search queries or secondary information giving inferences about sexual orientation, political and religious views, race, substance use, intelligence and overall personality, mental status, individual views and preferences (Kosinski, Bachrach, Stillwell, Kohli & Graepel, in press; Kosinski, Stillwell & Graepel, 2013).

As for the unprecedented possibilities to collect data, store big data and aggregate information that can be compared to big data Gestalt over time and society, privacy has leveraged into one of the most fragile areas of concern in the electronic age, demanding for legal protection, regulatory control and e-ethics (Flaherty, 1989). Today, the existing global privacy rights framework in the digital age has been criticized to be incoherent, inefficient and in need for revision. Global privacy protection shields are demanded to be established. Yet to this day, there is no economic framework on information sharing and privacy control. While – for instance – Posner (1981) criticizes privacy for concealing information, which reduces market efficiency; Lessig (2006) advocates for regulated online privacy. As of now we lack a behavioral decision making frame to explain the privacy paradox of the individual predicament between the humane-imbued preference to communicate and information share versus value of privacy. We have no behavioral economics description of inconsistencies and moderator variables in the decision between online information sharing behavior and retroactive preference reversal preferences in the eye of privacy concerns in the digital big data era.

A UTILITY THEORY OF INFORMATION SHARING AND PRIVACY

Building on classical utility theory, individuals are constantly evaluating competing choice options. Individuals weight alternative options based on their expected utility derived. Indifference curves would then connect points on a graph representing different quantities of two goods, between which an individual is indifferent.

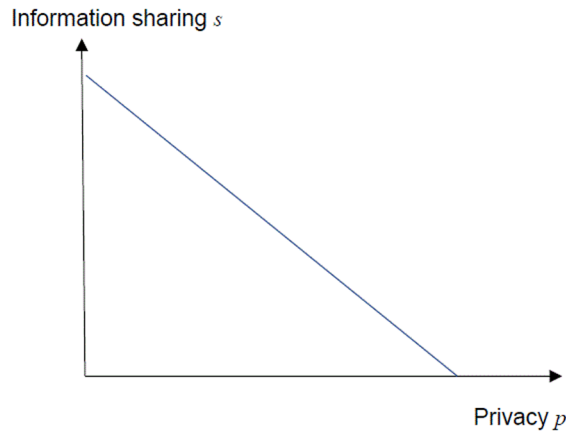
In the case of the privacy paradox of information sharing preferences and privacy values, a person would weights whether or not to share information s or choose the information to remain private p . The respective indifference curves would outline how much of information sharing s and privacy p can be enabled to end with the same utility given the budget of overall information held by the decision maker.

Graph 3 represents the respective indifference curves for information sharing s and privacy p . That is, the individual has no preference for one combination or bundle of information sharing or privacy over a different combination of the same curve. All points on the curve hold the same utility for the individual. The indifference curve is therefore the locus of various points of different combinations of privacy and information sharing providing equal utility to her or him. Indifference curves are thereby seen to represent potentially observable behavioral patterns for individuals over information bundles. The indifference curve for information sharing s and privacy p is subject to communication and information constraints, hence all information budgets and communication opportunities. There is only a finite amount of information and there may be environmental conditions determining if people can exchange and share information. As exhibited in graph 1, the indifference curve for information sharing s and privacy p is a straight line given the assumption that information sharing or privacy are substitutes.

Figure 1. Indifference curve (blue line) for information sharing s and privacy p given the total information and communication constraint

Information sharing s

Privacy p



While in classical economics, an individual was believed to always being able to rank any consumption bundles by order of preference (Jevons, 1871);² the indifference curve for information sharing s and privacy p subject to communication and information constraints may feature a hyper-hyperbolic element or temporal dimension. The information share moment may thereby be a reference point. At the moment of the information sharing decision, it may not be foreseeable what the future implication of the information sharing is.

In general, the costs and benefits of communication are assumed as linear subtraction of positive benefits of communication b_c minus the negative consequences of communication c_c . The nature of the problem is intertemporal as information sharers cannot foresee the future implications of their information sharing divided by variance σ (Prat, 2017).

$$\frac{b_c - c_c}{\sigma}. \quad (1)$$

However, the digital social media era has heralded a hyper-hyperbolic discounting fallibility. Individuals have lost oversight of the consequences of their individual information sharing given big data hoarding capabilities, which also allow drawing inferences about the individual in relation to others.

In the digital big data era, information share online may hold unforeseen risks of privacy merchants or social media capitalists that commercialize information reaping hidden benefits from the information provided (Etzioni, 2012; Puaschunder, 2017; The Economist, November 4, 2017). The subjective additive utility of information-shared tranche by tranche may underestimate the big data holder's advantage to reap benefits from information shared given unprecedented data storage and big data computation power advantages of the big data era. Unprecedented computational power and storage opportunities have created the possibility to hoard information over time and put it in context with the rest of the population in order to draw inferences about the information sharer (The New York Times, November 14, 2017). The digital age and era of instant information sharing have therefore heralded problems of individuals who give in their basic humane need for information communication to become vulnerable over time. The big data information holder may thereby benefit from the history of information and the relation of the individual's information in comparison to the general population to an unknown degree given missing e-literacy and transparency. Comparison to the general public may lead to an implicit underrepresentation and hence discrimination of vulnerable groups. For instance, certain groups that may not be represented online will therefore likely face an under-advocacy of their rights and needs.

While regular hyperbolic discounting captures a game theoretical predicament of the self now versus the self later, the information offering more of a Gestalt in the eyes of the big data holder, leverages hyperbolic discounting to a game theory against uncertainty on the end of the big data holder. The hyper-hyperbolic discounting fallibility therefore may describe that at the moment of information sharing, the individual has hardly any grasp what is implied in the giving up of privacy. The individual only focuses on the current moment trade-off between information sharing and privacy upholding, but hardly has any insights what the compiled information over time holds for big data moguls. As for holding computational and storage advantages, the social media big data moguls can form a *Gestalt*, which is more than the sheer sum of the individual information shared, also in comparison to the general populace's data. The shared information can also be resold to companies (Etzioni, 2012; The New York Times, November 14, 2017). In relation to other people's information, the big data moguls can make predictions about their choices and behaviors. Information can also be used for governance purposes, for instance tax compliance and border control mechanisms (Puaschunder, 2017). Some governments have recently used big data to check the accuracy of tax reports but also to detect people's political views when crossing borders (Puaschunder, 2017). Lastly, the use of big data inferences also implies hidden persuasion means – nudging can be turned against innocent information sharers who have no long-term and computational advantage to foresee the impact of the information share (The Economist, November 4, 2017; Puaschunder, 2017).

While behavioral economics hyperbolic discounting theory introduces the idea of time-inconsistency of preferences between an individual now and the same individual in the future; hyper-hyperbolic discounting underlines that in the case of information sharing preferences this fallibility is exacerbated since individuals lose control over their data and big data moguls can reap surplus value from the social media consumer-workers' information sharing and derive information compiled over time and in relation to the general norm to draw inferences about the innocent information sharer. With the modern digital era, all these features open an information sharer versus information reaper divide in the big data age (Puaschunder, 2017).

From the social media big data capitalist view, the information gain of one more person sharing information is exponentially rising. Hence, the marginal utility derived from one more person providing

information is increasing exponentially and disproportionately to the marginally declining costs arising from one more person being added to the already existing social media platform. Communication costs and benefits are assumed to not be additive and separable.

Expected Utility and Subjective Probability in the Digital Big Data Era

In accordance with neo-classical utility theory, decision makers weight alternatives based on the resulting consequences dependent on uncertain aspects of the environment. But in the digital big data era, individuals simply lack an oversight of the consequences of information sharing. Assumptions on preferences of information sharing are skewed leading to an underestimation of the consequences of amalgamated information and private information evaluated in relation to other's data. Assignment of utilities to the consequences are underinflated. The utility of information sharing is thus the underweighted sum of the utilities of the consequences.

Time Preferences

Following the standard neo-classical nomenclature of time preferences among the population, an information sharing preference over time is introduced. Multi-period decision-making addresses that for each time period, another set of preferences for the same options can be expected. The populace may therefore be theoretically categorized into

- (1) *Extreme impatience*: extreme information sharing as the individual values immediate pleasure of information sharing, information is shared without hesitation, impression management may play a role in this.
- (2) *Impatience*: discounting the future impact of information, uninformed information sharing nature. This is the case if an individual shares information although having a hunch that this information sharing may create problems in the future, called the privacy paradox.
- (3) *Eventual impatience*: discounting the future impact of information at some point in the future leads to controlled information sharing, very likely choosing what categories to expose to public.
- (4) *Time perspective*: related to hyper-hyperbolic discounting awareness, individuals may control information sharing, for instance, these individuals may participate in social media only to reap information from others but not contribute additional information beyond what is required. This type has a controlled privacy and is engaged in social media solely to reap benefits of other's information from social media networks.
- (5) *No time preference*: at the present time the individual neither discounts nor overcounts the future with respect to the present, which may be true for individuals who do not at all participate in social media communication and are blasé about information sharing and gaining information on social media,
- (6) *Persistence*: consistent preference structure regarding information sharing may result in informed information sharing with no regrets, and
- (7) *Variety*: consistently varying preference structure regarding information sharing, likely dependent on the content of information shared, may result in information sharing with regrets afterwards. These individuals have no stringent position towards information sharing or privacy preference, likely have categories for what to share and what not. This type varies in information preferences over time and by subject category.

This nomenclature addresses the problem of capturing valid preference structures over time that allow stringent predictions of choice behavior and strategies for future planning (Fishburn, 1968). The nomenclature also highlights that the selection of information sharing or privacy has an impact on our later choices. Addressing this predicament, Klein and Meckling (1958) suggest the best strategy given future uncertainties is to concentrate attention on immediate decisions that lead toward the main objective while preserving a reasonable degree of freedom in future choices. While Strotz (1957) considers the maximization of utility in an additive, discounted form over a continuous-time future and powerful research on hyperbolic discounting has unraveled pre-commitment and consistent planning as means to curb harmful decision-making fallibility. Yet the age of social media generated big data may impose novel hyperbolic discounting fallibility onto the information-sharing individual (Behears, Choi, Laibson, Madrian & Sakong, 2011; Chabris et al., 2008; Koopmans, 1964). Future research may test the reliability and validity of the nomenclature and unravel moderator variables and variances between different populations, e.g., such as age, cultural heritage, gender, etc.

Expected Utility and Subjective Probability

In accordance with neo-classical utility theory, alternatives are weighted based on the resulting consequences dependent on uncertain aspects of the environment. Assumptions on preferences between such alternatives lead to an assignment of utilities to the consequences and to the alternatives plus an assignment of subjective probabilities to the possible states of the environment. The utility of an alternative can therefore be written as a weighted sum of the utilities of the consequences. The weight for any alternative-consequence pair is the subjective probability associated with the states of the environment that yield the given consequence when the given alternative is used (Fishburn, 1968).

Regarding expected utility, the overall expected utility equation for information sharing and privacy reads

$$u = \sum w^* u_s + w^* u_p . \quad (2)$$

whereby w stands for weight, u_s is the utility of information sharing and u_p the utility of privacy.

The weighted expected utility equation reads

$$u(P) = P(x^1) u_s(x^1) + P(x^2) u_p(x^2) , \quad (3)$$

whereby $P(x^1) u_s(x^1)$ is the probability of information sharing utility and $P(x^2) u_p(x^2)$ the probability of privacy.

In the digital age, the utility of privacy is expected to have a marginal exponential value given the exponential rise of utility for the big data holder to reap benefits from data. The more data is held, the more complex relations can be unraveled by the big data holder. Information can be put into context of time and population correlates. While there is a marginally declining cost of an additional social media user using an established social network, there is a disproportionately large social network gain with another person joining for the social network provider, who can reap an excessive exponentially increasing marginal utility of another person joining and sharing another piece of information. Given the absence of any taxation of this gain, social media has leveraged into an IT monopoly (Soros, 2018).

FUTURE RESEARCH DIRECTIONS

The article presented a first theoretical introduction of a utility theory of information sharing and privacy. Potential limitations are that some communication may not be integrated in the framework, such as non-verbal communication or emotional responses. In general information exchange is very heterogeneous and vast international differences are assumed to exist. In addition, in what time communication online and under what circumstances decisions regarding communication and privacy are made, remains a completely undiscussed topic.

As a next research step, a stringent hypotheses testing of the presented problem is recommended. For instance, future research projects featuring a multi-methodological approach will help gain invaluable information about the actual performance and behavior regarding information sharing and privacy upholding. Interaction of individual on social media should be scrutinized in order to derive real-world relevant economic insights for legal and policy making purposes alongside advancing an upcoming scientific field. Following empirical investigations should employ a critical survey of the intersection of analytic and behavioral perspectives to decision making in information sharing. Literature discussion featuring a critical analysis how to improve e-literacy should be coupled with e-education and enhancement of e-ethicity. Research should be directed towards a critical analysis of the application of behavioral economics on hyper-hyperbolic discounting in the digital age. In the behavioral economics domain, both approaches, studying the negative implications of information sharing and decision making to uphold privacy but also finding ways how to train new media users wiser decisions should be explored. Interdisciplinary viewpoints and multi-method research approaches should be covered in the heterodox economics readings but also in a variety of independent individual research projects. Research support and guidance should be targeted at nurturing interdisciplinary research interests on privacy and information sharing in the fields of behavioral economics and public affairs.

More concretely, future studies should define the value that data has to individuals and data sovereignty in the international context. When people share information, they should be informed to consider what the benefit and value from sharing is for them and what the benefit for social media industrialists-capitalists is. The sovereignty of data and the human dignity of privacy should become debated as virtual virtue in the 21st century. Individuals should be informed that sharing data is a personal security risk, if considered to be asked for social media information upon entry of a country.

CONCLUSION

Overall, the presented piece can also serve as a first step towards advocating for education about information sharing in order to curb harmful information sharing discounting fallibility. From legal and governance perspectives, the outlined ideas may stimulate the e-privacy infringement regulations discourse in the pursuit of the greater goals of democratization of information, equality of communication surplus and upheld humane dignity and e-ethics in the big data era.

ACKNOWLEDGMENT

Financial support of the Academy of Behavioral Finance and Economics, American Academic Research Conference on Global Business, Economics, Finance and Social Sciences, Eugene Lang College of The

New School, Fee Board, Fritz Thyssen Foundation, George Washington University, Janeway Center Fellowship, New School for Social Research, Prize Fellowship, the Science and Technology Global Consortium, Research Association for Interdisciplinary Studies, the University of Vienna, Systema B, The New School Dean's Office, The New School Department of Economics, The New School Fee Board, The New School for Social Research, and Vernon Arts and Sciences is gratefully acknowledged. The author thanks Professor Andrea Prat for a most interesting lecture on 'Industrial Organization' at Columbia University during Fall 2017. All omissions, errors and misunderstandings in this piece are solely the author's.

REFERENCES

- Ainslie, G. (1992). *Picoeconomics: The interaction of successive motivational states within the person*. Cambridge University Press.
- Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting. In G. Loewenstein & I. Elster (Eds.), *Choice over time*. Russell Sage.
- Alberini, A., & Chiabai, A. (2007). Discount rates in risk versus money and money versus money trade-offs. *Risk Analysis*, 27(2), 483–498. doi:10.1111/j.1539-6924.2007.00899.x PMID:17511713
- Alchian, A. A. (1953). The meaning of utility measurement. *The American Economic Review*, 32, 26–50.
- Ali, N. S., & Benabou, R. (2017). *Image versus information: Changing societal norms and optimal privacy*. Cambridge, MA: National Bureau of Economic Research working paper 22203. Retrieved from https://scholar.princeton.edu/sites/default/files/rbenabou/files/imagevsinfo-21_002.pdf
- Allport, G. W. (1979). *The nature of prejudice*. Addison-Wesley.
- Altman, I. (1975). *The environment and social behavior: Privacy, personal space, territory, and crowding*. Brooks.
- Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica*, 19(4), 404–437. doi:10.2307/1907465
- Arrow, K. J. (1958). Utilities, attitudes, and choices: A review note. *Econometrica*, 26(1), 1–23. doi:10.2307/1907381
- Badger, G. J., Bickel, W. K., Giordano, L. A., Jacobs, E. A., Loewenstein, G., & Marsch, L. (2007). Altered states: The impact of immediate craving on the valuation of current and future opioids. *Journal of Health Economics*, 26(5), 865–876. doi:10.1016/j.jhealeco.2007.01.002 PMID:17287036
- Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it*. Princeton University Press. doi:10.1515/9781400837991
- Becker, G. M., & McClintock, C. G. (1967). Values: Behavioral decision theory. *Annual Review of Psychology*, 18(1), 239–286. doi:10.1146/annurev.ps.18.020167.001323 PMID:5333424
- Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, 96(4), 675–700. doi:10.1086/261558

Beshears, J. L., Choi, J. J., Laibson, D. I., Madrian, B. C., & Sakong, J. (2011). *Self-control and liquidity: How to design a commitment contract*. RAND Working Paper Series, WR- 895-SSA.

Besley, T., & Prat, A. (2006). Handcuffs for the grabbing hand? Media capture and government accountability. *The American Economic Review*, 96(3), 720–736. doi:10.1257/aer.96.3.720

Bosworth, R., Cameron, T. A., & DeShazo, J. R. (2006). *Preferences for preventative public health policies with jointly estimated rates of time preference*. School of Public Health and International Affairs: North Carolina State University.

Bowles, S. (2004). *Microeconomics: Behavior, institutions & evolution*. Princeton University Press.

Cairns, J., & van der Pol, M. (2008). Valuing future private and social benefits: The discounted utility model versus hyperbolic discounting models. *Journal of Economic Psychology*, 21(2), 191–205. doi:10.1016/S0167-4870(99)00042-2

Camerer, C. F., Loewenstein, G., & Rabin, M. (2004). *Advances in behavioral economics*. Princeton University Press. doi:10.1515/9781400829118

Cameron, T. A., & Gerdes, G. R. (2003). *Eliciting individual-specific discount rates*. Department of Economics: University of Oregon. doi:10.2139/ssrn.436524

Chabris, Ch., Laibson, D., & Schuldt, J. (2008). *Intertemporal choice*. Palgrave Dictionary of Economics.

Chapman, G. B. (1996). Expectations and preferences for sequences of health and money. *Organizational Behavior and Human Decision Processes*, 67(1), 59–75. doi:10.1006/obhd.1996.0065

Chung, S. H., & Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, 10(1), 67–74. doi:10.1901/jeab.1967.10-67 PMID:16811307

Colinsky, J. (1996). Why bounded rationality? *Journal of Economic Literature*, 34, 669–700.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, 2(2), 107–127. doi:10.1023/A:1009986005690

Cr mer, J., Garicano, L., & Prat, A. (2007). Language and the theory of the firm. *The Quarterly Journal of Economics*, 122(1), 373–407. doi:10.1162/qjec.122.1.373

DellaVigna, St., & Malmendier, U. (2004). Contract design and self-control: Theory and evidence. *The Quarterly Journal of Economics*, 119(2), 353–402. doi:10.1162/0033553041382111

DellaVigna, St., & Malmendier, U. (2006). Paying not to go to the gym. *The American Economic Review*, 96(3), 694–719. doi:10.1257/aer.96.3.694

DellaVigna, St., & Paserman, M. D. (2005). Job search and impatience. *Journal of Labor Economics*, 23(3), 527–588. doi:10.1086/430286

Delli Carpini, M. X., & Keeter, S. (1989). *What Americans know about politics and why it matters*. Yale University Press.

Doyle, J. R. (2013). Survey of time preference, delay discounting models. *Judgment and Decision Making*, 8(2), 116–135.

- Duflo, E., Banerjee, A., Glennerster, R., & Kothari, D. (2010). Improving immunization coverage in rural India: A clustered randomized controlled evaluation of immunization campaigns with and without incentives. *British Medical Journal*, 340(may17 1), 2220. doi:10.1136/bmj.c2220
- Duflo, E., Kremer, M., & Robinson, J. (2008). How high are rates of return to fertilizer? Evidence from field experiments in Kenya. *The American Economic Review*, 98(2), 482–488. doi:10.1257/aer.98.2.482
- Ebert, J. E., & Prelec, D. (2007). The fragility of time: Time-insensitivity and valuation of the near and far future. *Management Science*, 53(9), 1423–1438. doi:10.1287/mnsc.1060.0671
- Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, 51(4), 380–417. doi:10.1037/h0053870 PMID:13177802
- Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, 12(1), 473–498. doi:10.1146/annurev.ps.12.020161.002353 PMID:13725822
- Epper, T., Fehr-Duda, H., & Bruhin, A. (2011). Viewing the future through a warped lens: Why uncertainty generates hyperbolic discounting. *Journal of Risk and Uncertainty*, 43(3), 169–203. doi:10.1007/11166-011-9129-x
- Estle, S. J., Green, L., Myerson, J., & Holt, D. D. (2007). Discounting of money and directly consumable rewards. *Psychological Science*, 18(1), 58–63. doi:10.1111/j.1467-9280.2007.01849.x PMID:17362379
- Etzioni, A. (2012). The privacy merchants: What is to be done? *The Journal of Constitutional Law*, 14(4), 929–951.
- Fishburn, P. C. (1968). Utility theory. *Management Science*, 14(5), 335–378. doi:10.1287/mnsc.14.5.335
- Flaherty, D. (1989). *Protecting privacy in surveillance societies: The federal republic of Germany, Sweden, France, Canada, and the United States*. The University of North Carolina Press.
- Fowler, R., Hodge, B., Kress, G., & Trew, T. (1979). *Language and control*. Routledge.
- Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2), 351–401. doi:10.1257/jel.40.2.351
- Gaudeul, A., & Giannetti, C. (2017). The effect of privacy concerns and social network formation. *Journal of Economic Behavior & Organization*, 141, 233–253. doi:10.1016/j.jebo.2017.07.007
- Gentner, D. (2002). The psychology of mental models. In N. J. Smelser & P. B. Bates (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 9683–9687). Elsevier.
- Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison. *Psychological Science*, 5(1), 33–36. doi:10.1111/j.1467-9280.1994.tb00610.x
- Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130(5), 769–792. doi:10.1037/0033-2909.130.5.769 PMID:15367080
- Hansen, A. (2006). Do declining discount rates lead to time inconsistent economic advice? *Ecological Economics*, 60(1), 138–144. doi:10.1016/j.ecolecon.2006.03.007
- Harrison, G. W., Lau, M. I., & Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *The American Economic Review*, 92(5), 1606–1617. doi:10.1257/000282802762024674

- Henderson, N., & Bateman, I. (1995). Empirical and public choice evidence for hyperbolic social discounting rates and the implications for intergenerational discounting. *Environmental and Resource Economics*, 5(4), 413–423. doi:10.1007/BF00691577
- Heyman, G. M. (1996). Resolving the contradictions of addiction. *Behavioral and Brain Sciences*, 19(4), 591–610. doi:10.1017/S0140525X00042990
- Hoch, St., & Loewenstein, G. (1991). Time-inconsistent preferences and consumer self-control. *The Journal of Consumer Research*, 17(4), 492–507. doi:10.1086/208573
- Hornsby, J. (2007). *An empirical investigation of the effects of discounting on privacy related decisions* (Dissertation). Nova Southeastern University.
- Horowitz, J. K., & Carson, R. T. (1990). Discounting statistical lives. *Journal of Risk and Uncertainty*, 3(4), 403–413. doi:10.1007/BF00353349
- Iyengar, S., & Kinder, D. R. (1987). *News that matter*. University of Chicago Press.
- Johnson, D. (2009). *Ethical theory and business*. Pearson Prentice Hall.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristic and biases*. Cambridge University Press. doi:10.1017/CBO9780511809477
- Kahneman, D., & Thaler, R. (1991). Economic analysis and the psychology of utility: Applications to compensation policy. *The American Economic Review*, 81, 341–346.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. doi:10.2307/1914185
- Keeler, E. G., & Cretin, S. (1983). Discounting of life-saving and other non-monetary effects. *Management Science*, 29(3), 300–306. doi:10.1287/mnsc.29.3.300
- Keller, L.R. & Strazzer, E. (2002). Examining predictive accuracy among discounting models. *Journal of Risk and Uncertainty*, 24(2), 143–160.
- Kirby, K. N. (1997). Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology. General*, 126(1), 54–70. doi:10.1037/0096-3445.126.1.54
- Kirby, K. N., & Herrnstein, R. J. (1995). Preference reversals due to myopic discounting of delayed reward. *Psychological Science*, 6(2), 83–89. doi:10.1111/j.1467-9280.1995.tb00311.x
- Kirby, K. N., & Marakovic, N. N. (1995). Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes*, 64(1), 22–30. doi:10.1006/obhd.1995.1086
- Klein, B., & Meckling, W. (1958). Application of operations research to development decisions. *Operations Research*, 6(3), 352–363. doi:10.1287/opre.6.3.352
- Koopmans, T. C. (1964). On flexibility of future preferences. In M. W. Shelly & G. L. Bryan (Eds.), *Human Judgments and Optimality*. Wiley.

Kosinski, M., Bachrach, Y., Stillwell, D. J., Kohli, P., & Graepel, T. (in press). Manifestations of user personality in website choice and behavior on online social networks. *Machine Learning*.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), 5802–5805. doi:10.1073/pnas.1218772110 PMID:23479631

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2), 443–477. doi:10.1162/003355397555253

Laibson, D., Repetto, A., & Tobacman, J. (2003). A debt puzzle. In P. Aghion, R. Frydman, J. Stiglitz, & M. Woodford (Eds.), *Knowledge, information, and expectations in modern economics: In honor of Edmund S. Phelps* (pp. 228–266). Princeton University Press.

Laux, F. L., & Peck, R. M. (2007). *Economic perspectives on addiction: Hyperbolic discounting and internalities*. Social Science Research Network working paper. Retrieved November 3, 2015 at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1077613

Lessig, L. (2006). *Code: Version 2.0*. Basic.

Loewenstein, G. (1992). The fall and rise of psychological explanations in the economics of intertemporal choice. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 3–34). Sage.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review*, 100(1), 91–108. doi:10.1037/0033-295X.100.1.91

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. Dover.

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R.D. Luce, R.R., Bush & E. Galanter (Eds.), *Handbook of mathematical psychology*, (pp. 103-189). New York: Wiley.

Madden, G. L., Bickel, W. K., & Jacobs, E. A. (1999). Discounting of delayed rewards in opioid-dependent outpatients: Exponential or hyperbolic discounting functions? *Experimental and Clinical Psychopharmacology*, 7(3), 284–293. doi:10.1037/1064-1297.7.3.284 PMID:10472517

Majumdar, T. (1958). *The measurement of utility*. Macmillan.

Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. *Econometrica*, 18(2), 111–141. doi:10.2307/1907264

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. Oxford University Press.

Mastrobuoni, G., & Weinberg, M. (2009). Heterogeneity in intra-monthly consumption patterns, self-control, and savings at retirement. *American Economic Journal. Economic Policy*, 1(2), 163–189. doi:10.1257/pol.1.2.163

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. M. Commons, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior*. Lawrence Erlbaum.

McClure, S., Ericson, K., Laibson, D., Loewenstein, G., & Cohen, J. (2007). Time discounting for primary rewards. *The Journal of Neuroscience*, 27(21), 5796–5804. doi:10.1523/JNEUROSCI.4246-06.2007 PMID:17522323

Meier, St., & Sprenger, Ch. (2010). *Stability of time preferences*. IZA Discussion Paper 4756, Institute for the Study of Labor.

Meyer, A. G. (2009). *Estimating individual level discount factors and testing competing discounting hypotheses* (Dissertation). University of Colorado.

Meyer, A. G. (2013). Estimating discount factors for public and private goods and testing competing discounting hypotheses. *Journal of Risk and Uncertainty*, 46(2), 133–173. doi:10.1007/11166-013-9163-y

Milkman, K. L., Rogers, T., & Bazerman, M. H. (2008). Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psychological Science*, 3(4), 324–338. doi:10.1111/j.1745-6924.2008.00083.x PMID:26158952

Milkman, K. L., Rogers, T., & Bazerman, M. H. (2009). Highbrow films gather dust: Time- inconsistent preferences and online DVD rentals. *Management Science*, 55(6), 1047–1059. doi:10.1287/mnsc.1080.0994

Murphy, J. G., Vuchinich, R. E., & Simpson, C. A. (2001). Delayed reward and cost discounting. *The Psychological Record*, 51, 571–588.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, 64(3), 263–276. doi:10.1901/jeab.1995.64-263 PMID:16812772

Neuman, W., Russel, M. R. J., & Crigler, A. N. (1992). *Common knowledge*. Chicago University Press. doi:10.7208/chicago/9780226161174.001.0001

Norris, P., & Sanders, D. (2003). Message or medium? Campaign learning during the 2001 British General Election. *Political Communication*, 20(3), 233–262. doi:10.1080/10584600390218878

Okada, E. M., & Hoch, S. J. (2004). Spending time versus spending money. *The Journal of Consumer Research*, 31(2), 313–323. doi:10.1086/422110

Olken, B. A. (2009). Do television and radio destroy social capital? Evidence from Indonesian villages. *American Economic Journal. Applied Economics*, 1(4), 1–33. doi:10.1257/app.1.4.1

Orwell, G. (1949). 1984. New York: Harcourt Brace.

Oster, S.M. & Scott-Morton, F.M.S. (2005). Behavioral biases meet the market: The case of magazine subscription prices. *Advances in Economic Analysis & Policy*, 5(1), 1-32.

Oulasvirta, A., Suomalainen, T., Hamari, J., Lampinen, A., & Karvonen, K. (2014). Transparency of intentions decreases privacy concerns in ubiquitous surveillance. *Cyberpsychology, Behavior, and Social Networking*, 17(10), 633–638. doi:10.1089/cyber.2013.0585 PMID:25226054

Petry, N. M., & Casarella, T. (1999). Excessive discounting of delayed rewards in substance abusers with gambling problems. *Drug and Alcohol Dependence*, 56(1), 25–32. doi:10.1016/S0376-8716(99)00010-1 PMID:10462089

Posner, R. A. (1981). The economic of privacy. *The American Economic Review*, 71(2), 405–409.

Prat, A. (2005). The wrong kind of transparency. *The American Economic Review*, 95(3), 862–877. doi:10.1257/0002828054201297

- Prat, A. (2017). *Lecture notes 'Industrial Organization'*. Columbia University.
- Prat, A., & Strömberg, D. (2005). *Commercial television and voter information*. CEPR Discussion Paper 4989.
- Prat, A., & Strömberg, D. (2013). *The political economy of mass media*. New York: Columbia University working paper. Retrieval at <http://www.columbia.edu/~ap3116/papers/mediasurvey11.pdf>
- Puaschunder, J. M. (2017). Nudgital: Critique of Behavioral Political Economy. *Archives of Business Research*, 5(9), 54–76. doi:10.14738/abr.59.3623
- Puaschunder, J. M., & Schwarz, G. (2012). *The future is now: How joint decision making curbs hyperbolic discounting but blurs social responsibility in the intergenerational equity public policy domain*. Harvard University Situationist Law and Mind Sciences Working Paper.
- Putnam, H. (2002). On the rationality of preferences. In H. Putnam (Ed.), *The Collapse of the Fact-Value Dichotomy and Other Essays*. Harvard University Press.
- Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. Simon & Schuster.
- Rachlin, H., Raineri, A., & Cross, D. (1991). Subjective probability and delay. *Journal of the Experimental Analysis of Behavior*, 55(2), 233–244. doi:10.1901/jeab.1991.55-233 PMID:2037827
- Read, D. (2001). Is time-discounting hyperbolic or subadditive? *Journal of Risk and Uncertainty*, 23(1), 1, 5–32. doi:10.1023/A:1011198414683
- Read, D., Frederick, S., & Airoldi, M. (2012). Four days later in Cincinnati: Longitudinal tests of intertemporal preference reversals due to hyperbolic discounting. *Acta Psychologica*, 140(2), 177–185. doi:10.1016/j.actpsy.2012.02.010 PMID:22634266
- Read, D., Loewenstein, G., & Kalyanaraman, S. (1999). Mixing virtue and vice: The combined effects of hyperbolic discounting and diversification. *Journal of Behavioral Decision Making*, 12, 257–273. doi:10.1002/(SICI)1099-0771(199912)12:4<257::AID-BDM327>3.0.CO;2-6
- Read, D. R., & van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, 76(2), 189–205. doi:10.1006/obhd.1998.2803 PMID:9831521
- Reed, D. D., & Martens, B. K. (2011). Temporal discounting predicts student responsiveness to exchange delays in a classroom token system. *Journal of Applied Behavior Analysis*, 44(1), 1–18. doi:10.1901/jaba.2011.44-1 PMID:21541113
- Reuben, E., Sapienza, P., & Zingales, L. (2010). Time-discounting for primary and monetary rewards. *Economics Letters*, 106(2), 125–127. doi:10.1016/j.econlet.2009.10.020
- Salanié, F., & Treich, N. (2005). Over-savings and hyperbolic discounting. *European Economic Review*, 50(6), 1557–1570. doi:10.1016/j.eurocorev.2005.05.003
- Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies*, 4(2), 155–161. doi:10.2307/2967612
- Sen, A. K. (1971). Choice functions and revealed preference. *The Review of Economic Studies*, 38(3), 307–317. doi:10.2307/2296384

- Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, 6(4), 317–344.
- Sen, A. K. (1993). Internal consistency of choice. *Econometrica*, 61(3), 495–521. doi:10.2307/2951715
- Sen, A. K. (1995). Rationality and social choice. *The American Economic Review*, 85(1), 1–24.
- Sen, A. K. (1997). Maximization and the act of choice. *Econometrica*, 65(4), 745–780. doi:10.2307/2171939
- Sen, A. K. (2002a). Consistency of choice. In A. Sen (Ed.), *Rationality and freedom*. Harvard University Press.
- Sen, A. K. (2002b). Goals, commitment, and identity. In A. Sen (Ed.), *Rationality and freedom*. Harvard University Press.
- Shackle, G. L. S. (1955). *Uncertainty in economics and other reflections*. Cambridge University Press.
- Shah, A. K., & Oppenheimer, D. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134(2), 207–222. doi:10.1037/0033-2909.134.2.207 PMID:18298269
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x
- Shui, H., & Ausubel, L. M. (2004). *Time inconsistency in the credit card market*. Social Science Research Network working paper. Retrieved October 25, 2015 at <http://ssrn.com/abstract=586622>
- Simon, H. A. (1959). Theories of decision making in economics and behavioral science. *The American Economic Review*, 49, 253–283.
- Slonim, R., Carlson, J., & Bettinger, E. (2007). Possession and discounting behavior. *Economics Letters*, 97(3), 215–221. doi:10.1016/j.econlet.2007.03.018
- Snyder, J. M. Jr, & Strömberg, D. (2010). Press coverage and political accountability. *Journal of Political Economy*, 118(2), 355–408. doi:10.1086/652903
- Solove, D. J. (2008). *Understanding privacy*. Harvard University Press.
- Soros, G. (2018). *Remarks delivered at the World Economic Forum in Davos*. Retrievable at <https://www.georgesoros.com/2018/01/25/remarks-delivered-at-the-world-economic-forum/>
- St. Jevons, W. (1871). *The theory of political economy*. Macmillan. Retrieved from <http://www.econlib.org/library/YPDBooks/Jevons/jvnPECover.html>
- Sterner, T. (1994). Discounting in a world of limited growth. *Environmental and Resource Economics*, 4(5), 527–534. doi:10.1007/BF00691927
- Strotz, R. H. (1957). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3), 165–180. doi:10.2307/2295722
- Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, 8(3), 201–207. doi:10.1016/0165-1765(81)90067-7
- Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, 89(2), 392–406. doi:10.1086/260971

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge. Improving decisions about health, wealth, and happiness*. Yale University Press.

Trope, Y., & Fishbach, A. (2000). Counteractive self-control in overcoming temptation. *Journal of Personality and Social Psychology*, 79(4), 493–506. doi:10.1037/0022-3514.79.4.493 PMID:11045735

van der Pol, M., & Cairns, J. (2001). Estimating time preferences for health using discrete choice experiments. *Social Science & Medicine*, 52(9), 1459–1470. doi:10.1016/S0277-9536(00)00256-2 PMID:11286369

von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton University Press.

Warner, J. T., & Pleeter, S. (2007). The personal discount rate: Evidence from military downsizing programs. *The American Economic Review*, 91(1), 33–53. doi:10.1257/aer.91.1.33

Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard Law Review*, 4(5), 193–220. doi:10.2307/1321160

Wertenbroch, K. (1998). Consumption self-control by rationing purchase quantities of virtue and vice. *Marketing Science*, 17(4), 317–337. doi:10.1287/mksc.17.4.317

Witzany, G. (2012). *Biocommunication of fungi*. Springer. doi:10.1007/978-94-007-4264-2

Zauberman, G., Kim, B. K., Malkoc, S. A., & Bettman, J. R. (2009). Discounting time and time discounting: Subjective time perception and intertemporal preferences. *JMR, Journal of Marketing Research*, 46(8), 543–556. doi:10.1509/jmkr.46.4.543

ADDITIONAL READING

Beerbaum, D., & Puauschunder, J. M. (2018). A behavioral economics approach to digitalisation: The case of a principles-based taxonomy. Social Science Research Network working paper retrievable at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3210207

Puaschunder, J. M. (2017). Nudging in the digital big data era. *European Journal of Economics. Law and Politics*, 4(4), 18–23.

Puaschunder, J. M. (2017). Nudgital: Critique of Behavioral Political Economy. *Archives of Business Research*, 5(9), 54–76. doi:10.14738/abr.59.3623

Puaschunder, J.M. (2017). The nudging divide in the digital big data era. *International Journal of Research in Business, Economics and Management*, 4, 11-12, 49-53.

Puaschunder, J. M., & Schwarz, G. (2012). *The future is now: How joint decision making curbs hyperbolic discounting but blurs social responsibility in the intergenerational equity public policy domain*. Harvard University Situationist Law and Mind Sciences Working Paper.

KEY TERMS AND DEFINITIONS

Information Sharing: Implying giving up privacy is at the core of communication. Communication can be verbal and non-verbal. Comprising very many different domains ranging from business, politics, interpersonal, social to mass media; communication is a humane-imbued wish and center core of every functioning society.

Privacy: Is a core human value as people choose what information to share with whom and like to protect some parts of their selves in secrecy. Protecting people's privacy is a codified virtue around the world to uphold the individual's dignity.

Privacy Laws: In many countries have started to adapt to changes in technology in order to cope with unprecedented constant information surveillance possibilities, big data storage opportunities and computational power peaks.

Social Media Tools: Have become large-scale factories with unpaid labor for information sharing. The accuracy of this information also appears questionable, with about 83.09 million accounts assumed to be fake.

Utility Theory: Is concerned with people's choices and decisions based on preferences and values.

ENDNOTES

¹ e.g., Asian-Pacific Economic Cooperation, Australia, Brazil, Canada, China, European Union, Italy, Japan, Korea, Organisation for Economic Co-operation and Development, South Africa, United Kingdom, United Nations, United States, Universal Declaration of Human Rights – to name a few.

² <http://www.econlib.org/library/YPDBooks/Jevons/jvnPE.html>

The 2018 Facebook Data Controversy and Technological Alienation

4

Ananda Mitra

Wake Forest University, USA

Yasmine Khosrowshahi

Wake Forest University, USA

INTRODUCTION

The controversy surrounding the way in which data was obtained and used by a private corporation, Cambridge Analytica, in connection with the US Presidential election of 2016 begs a series of questions and points to the critical connection between Big Data and a variety of fundamental human imperatives such as privacy and freedom of expression amongst others (Cadwalladr and Graham-Harrison, 2018). This chapter uses the issues surrounding the controversy as a point of departure to explore two related issues – first the way in which the “technologically alienated” individual makes technological decisions without being mindful of the consequences of the technological choices and secondly the way in which the alienation can be utilized to obtain detailed information from the structured and unstructured data that makes up the gigantic corpus of Big Data.

The key to the controversy was the fact that a corporation was able to collect information from individuals in a manner where the people of whom data was collected felt that they were unaware of the fact that their data was being collected. What happens after that is important, but the goal of this essay is to examine the claim by individuals, and their representatives that the data collection happened in a surreptitious manner. While, this essay is not meant to condone what happened, but it is to offer an alternative argument that places a great deal of responsibility on individual users of digital systems. It is useful to consider the notion of alienation in this context.

BACKGROUND

Alienation From Technology

The notion of alienation has been a key philosophical construct in thinking of people in societies. The notion of alienation has been a focus of study from many different perspectives. For example, the approaches taken by political economist Karl Marx and by the psychiatrist Sigmund Freud have been some of the more well-known perspectives on the idea of alienation. In the works of Marx¹ and Freud² there was the common theme that alienation involved the distancing between two entities. Thus, Marx argued that the practice of capitalism produces a distance between the laborer and the product of his or her labor. Specifically, Marx was concerned about the way in which the laborer is separated from the profit that is gleaned from the work of the laborer. Working in a time period similar to Marx, Freud, on the other hand, took a different approach to alienation because not only was Freud not interested in a political

DOI: 10.4018/978-1-7998-3479-3.ch032

economic theory but also there was no great urgency in Freud's analysis on reducing alienation. For Freud, alienation was essentially the natural split between the conscious and the subconscious aspects of the human mind, and from the psychoanalytic perspective of alienation, was a part of the evolutionary process of being human and there was no specific question of overcoming this alienation. To Freud, humans were expected to live with a certain degree of alienation which was internal to the human being and an essential part of life. While Marx and Freud have been the two key scholars who have historically been connected with the idea of alienation, others like Erich Fromm³ and Emile Durkheim⁴ have also used and developed notions of alienation to underscore the distance between the human being and society. In most cases the focus has been on identifying the factors that cause the alienation and then finding mechanisms for reducing the alienation to an acceptable level. Eventually, the notion of alienation actually remains a very well-examined notion in psychology, philosophy, politics, and related disciplines and in most cases the emphasis has been on the ways in which the human being is being distanced from events around him or her. In much of the discussions about alienation, thinkers have tried to find ways to reduce the extent of alienation. The desire to reduce the sense of distance between the individual and the surroundings, particularly when alienation is triggered by external forces, is motivated by the fact that the outcome of alienation can be fundamentally painful to the individual. To feel that one is distanced from the surrounding, or to feel unfamiliar and unacknowledged can be a debilitating experience. The sense of loss of control that can emanate being disempowered has been the focus of discussions in politics and economy for a long time and attempts have been made to understand why alienation occurs and how it can be overcome. No wonder, alienation has received so much attention because it is considered to be an unpleasant and abnormal condition because of the uncertainty and anxiety it can produce. Most of the scholars would argue that to be alienated is essentially an uncomfortable feeling.

While this brief review of the understanding of alienation draws from the traditional understanding of the notion of alienation, I use and expand on the idea of alienation to primarily refer to the idea of distancing (Mitra, 2010). To me the notion of alienation is one that produces a distance between the human being and a specific phenomenon. Specifically, the notion of alienation when considered with respect to technology can be considered as follows:

In most cases, the more sophisticated the individual thinks the technology is, the greater is the alienation of the individual from the technology. Within this relationship between technology and alienation the idea of sophistication is two-pronged. On one hand is the measurable sophistication of a tool going, for instance, from the primitive bow and arrow to the Kalashnikov, and on the other hand is the perceived sophistication of the technology, in terms of how the user thinks about the technology. The latter refers to what the individual thinks about a tool often with limited understanding of the tool. Both these aspects of sophistication play a part in the process of alienation but often operate in different ways since the loci of these two forms of sophistication are in different places – in the technology and in the individual. It can be argued that the perception of sophistication is partly based in the level of knowledge the individual has about a specific technology. A technology can appear sophisticated simply because the person using does not know much about the technology and that ignorance creates the aura of heightened sophistication and increased alienation (Mitra, 2012).

The distance can remain unrecognizable and thus unexplored unless an explicit attempt is made to understand the nature of the alienation. Thus, in the analysis that follows in this essay, I am interested in tracing the nature of alienation between people and the technologies that surround them. In the cur-

rent analysis, the focus is on the way the alienated user remained alienated from a tool just because they perceived it to be sophisticated leading to a set of issues that eventually affected the user.

The Situation

The specific situation under consideration came to the public view in the spring of 2018 when it was reported in several media outlets that a private corporation called Cambridge Analytica was able to access data of Facebook users and use that data for many different political purposes that may have affected the outcome of the Presidential election of 2016 in America. It is estimated that nearly 87 million users were affected by the scandal. Staff writers of Forbes magazine explained that the scandal actually started several years back when Facebook exposed the data from millions of users to an academic researcher (DeMars, 2018). That academics researcher, a Russian-American named Aleksandr Kogan, used a quiz app, similar to a personality quiz app Psychometrics Centre made, called “thisisyourdigitallife” (Mer-edith, 2018). The app was used to pull more data from users than they should have been allowed to. The Atlantic Media Company says that 270,000 people installed Kogan’s app on their Facebook account. Just like any Facebook developer then, Kogan was able to access data about those users and their friends (Meyer, 2018). Kogan then sold all the data to Cambridge Analytica (CA) to then use as part of the 2016 US election for marketing and advertising for candidate Trump. Numerous news outlets including the *Al Jazeera*, the *London Observer* and *The New York Times* reported that Cambridge Analytica used the personal information of millions of users to build software to target potential swing voters in political campaigns including in the 2016 election for candidate Donald Trump. The information created messages targeted at users on Facebook. Christopher Wylie, a key player in the scandal, says, “We exploited Facebook to harvest millions of people’s profiles and built models to exploit what we knew about them and target their inner demons. That was the basis the entire company was built on” (Staff, *Al Jazeera*, 2018).

The Analysis

To demonstrate the issue of alienation the first step was to gain a better understanding of the way in which the general public, the users of digital networking systems, interpreted the situation and responded to the situation. The specific method used in the analysis was to create a narrative about the situation from the perspective of everyday users. This approach relies on the narrative paradigm and utilizes narrative analysis to discover the narrative, or the story, that emerges within the users.

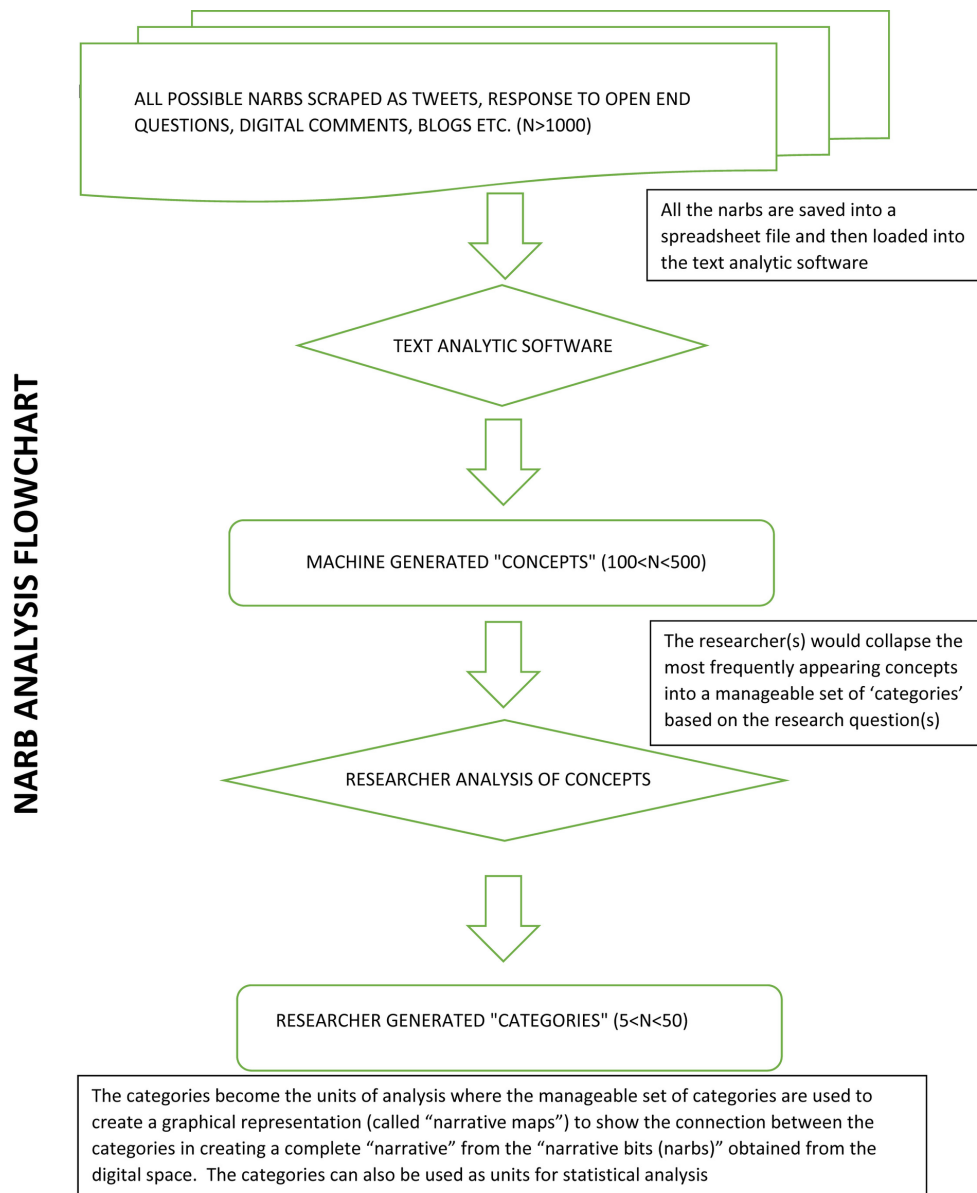
The construct of looking at culture, communication and human behavior and beliefs through the lens of a narrative was suggested by the work of Walter Fisher who in turn based his work on earlier scholars who identified the importance of story-telling and narratives in the process of making meaning of everyday life as people operate within the cultural and social spaces that they occupy (see, e.g., Andrews, 1982; MacIntyre, 1981; Ricouer, 1977, 1983, 1984; White, 1984). Fisher suggested, people can operate as a narrative being where the act of creating and articulating a coherent and rational narrative becomes a part of being human. In doing this, the narrative paradigm offers a unique perspective on the way of analyzing and understanding communication, and the way in which people act.

There are many aspects of the narrative paradigm that merit examination in the early Twenty-first Century of Big Data, particularly when the analysis of large data sets seeks to provide an account of how persons come to believe and behave (see, e.g., Kosinski, et. al. 2013).

The process of narrative analysis begins by collecting a corpus of narrative bits – narbs - that can be used to create a narrative and understand the story that emerges within the community of users rep-

resented by the narbs. Very soon after the news about Facebook became public, a topic (hashtag) was created on the micro-blogging site, Twitter, and many users tweeted with the DeleteFacebook hashtag to promote the end of using the digital network platform (Rosenberg, 2018). For the purpose of this essay, the narbs were composed of nearly 11,000 tweets that were collected for that specific hashtag. The tweets were collected, and the content of the tweets were then analyzed to extract a narrative map which is a visual representation of a narrative. In the analysis offered here, the automated text analysis using the SPSS Text Analytics tool is coupled with contextualizing of the dictionary and offering an intermediary step that mimics the traditional coding process of earlier textual analysis allowing for the recognition and incorporation of the context under which the unstructured big data has been produced, as demonstrated in Figure 1.

Figure 1. Narrative map of the categories extracted from the narbs



The Story

The narrative codes also explain the relationship among categories that were spoken about with a negative affect directly linked to it on the map. The triangle for users, Facebook, and negative affect validates the point that users had a negative affect towards Facebook after the scandal broke out. The lines show a strong relationship between users and negative affect and user and Facebook but a weak relationship between Facebook and negative affect. This indicates that the focus is around the users more than Facebook. Many sources discuss what Facebook did to users, not stating a negative or positive affect but rather just an understanding of how it took users information. This fits the triangle and if there are sources written about this relationship then many of the tweets extracted must have done the same. It is also clear that the triangles that negative affect create are all connected with users as well. Therefore, Internet bill of rights, data, Ad Revenue, and profit all create triangles individually with user and negative affect to show that users are negatively affected by each of these categories. Here again, the users, whose tweets are the basis of the narrative map, deflect attention from themselves to everything else around the use of the Facebook platform without concern for the fact that the alienation of the user also plays a role in the event that occurred.

Negative affect seems to be the category with the most relationships and interconnectedness on the map. The strongest of which is the triangle for FB trackers, attentiveness, and negative affect discussed above. What is also clear in the narrative is the fact that there is no connection between negative affect and Zuckerberg, which suggests that, in this narrative, there is no negative affect for Zuckerberg. While many categories had negative affects however, many categories had an equal amount of positive and negative affects. There were also many other categories such as users, Facebook, privacy, and attentiveness that had both positive and negative affects indicating that for many categories, there were split feelings and opinions.

In addition to the thickness of the lines indicating a stronger relationship, the size of the circle also indicates the number of narbs the category has. For instance, negative affect has the largest circle on the map, meaning there are the most narbs for that category. Aside from negative affect, Facebook Tracker, attentiveness, and Facebook all have larger circles than the rest of the categories and these are the categories talked about most. Most of the categories on the map have the same size circle and same line thickness which means that they are equally all talked about in regard to the hashtag. It is clear that negative affect and its relationship with the categories, especially Facebook Tracker and attentiveness, are the most spoken about.

Another interesting triangle is the connection with users, Zuckerberg, and Facebook. Since the Facebook scandal, many users have blamed the founder of Facebook, Mark Zuckerberg, thus it would make sense that there would be a correlation among the three categories. In addition, since Zuckerberg went in front of the American Congress, there was even more talk about him in relation to users and Facebook. It even seems as if the lines should be thicker from much discussion about the relationship among these three categories. Many articles and sources have surfaced regarding the situation in front of congress. Facebook even streamed it live however this narrative indicates that there was not a strong connection with Zuckerberg, Facebook, and users. Eventually the question that is most important in the context of this essay is the degree to which the users were alienated from the technology, and how that alienation has an effect on the way responsibility is attributed. In other words, “who is to blame” for this event.

Who to Blame?

An important aspect of the narrative is the shifting of the blame for the incident to outside entities, especially the Facebook application and its developer and promoter. As indicated in the story, the users

do not consider the fact that there was a lack of understanding, and therefore an alienation, on the part of the user. There is a strong negative affect about the process of being tracked, and a negative opinion about the need to be attentive to the process of being tracked. It is only after the fact that the scandal happened, when the data was collected, that there appears to be a strong connection between the need to be attentive and the annoyance about that need. This is what is considered to be alienation, where the user gets distant from the tool, continues to use the tool, and it nearly disappears from sight, i.e., becomes ubiquitous (Weiser, 1991). At that point, the user is so alienated from the tool, that the user is inattentive and expects that the tool is working as the user wants it to work. It is only, when the tool “breaks” that the user suddenly becomes aware of the tool, and its accessories, such as the “FB Trackers,” and report a negative affect about it.

The fact remains that the trackers have been around for a long period of time, and will continue to be around in the future, and the alienated user will again become inattentive until the next crisis happens. It is precisely this alienation, and consequent inattentiveness, that allows systems such as Facebook collect data about the alienated user. The example of the event from the Facebook crisis drew attention to the fact that data is collected pervasively and constantly by the digital systems that the users continue to use. Indeed, it can be argued that the growth in the amount of data – and the consequent focus on “Big Data” – is voluminous, grows at great velocity and contains highly granulated details. Yet, the users are alienated from the process of data collection because of the process has receded from sight, and happens “under the hood,” and the users continue to provide the data so that the overall “engine” of the essential digital systems continue to benefit the user. Just as a user can remain alienated from the exact way in which a car operates, but can effectively drive the car, the users of the digital systems can effectively navigate around an online banking system while being alienated from the process that they are using to access the system and the amount of data that they are providing the system. Only when that data is breached does the user become concerned, and as demonstrated in this essay, shifts the blame to the data collection and management process without acknowledging the intrinsic responsibility of the user to be mindful of the way in which the technology operates.

A final crucial aspect of the combination of the alienation and increase in user-generated data that is available for use as in the case of the Facebook incident is the fact that the everyday material practices, from purchasing a bar of soap from a distribution outlet that sells soap from a virtual shop, to playing a game provided by digital social networking platform such as Facebook, have become intimately integrated into the fabric of “everyday life” for most people in the developed and developing World. There is an inescapable pervasiveness to this process that set the foundation for the Facebook crisis. It is only a matter of time before other such critical matters happen where the need for convenience and to increase the efficiency of processes ubiquitous tools are deployed which can eventually be compromised. For instance, one of the largest digital/social experiment has been the creation of a biometric database of over one billion people in India. The system allowed for great conveniences – from opening a bank account to obtaining a cell phone connection – all with a finger print scanner. Yet, the question will have to be settled how the narrative would develop and circulate following reports that the data has been compromised (Khaira, et. al., 2018; Whittaker, 2018). Consequently, using the experience of the Facebook crisis, and anticipating similar events in the future, it is useful to consider some ways in which the user must negotiate the essential terrain of existing in a system where the “real” person is already and always converted as a part of Big Data.

SOLUTIONS AND RECOMMENDATIONS

The emerging data environment and the issues highlighted here, first requires a reconsideration of the tension between the need for mindfulness of one's data-related actions and the expectation that one has the freedom to "speak" and the speech may not be misused. The data suggests that among those who belong to the group who tweeted there was a negative affect towards the need for being attentive. The underlying theme is that one should not have to be continuously mindful about every data-related action. Indeed, the data suggests that users feel that the transparency of the technology should encourage a sense of alienation, and that alienation should not be a cause for concern. Having to be mindful of data-related actions is not considered to be a comfortable condition. For example, users of search engines should not have to be concerned about what they are searching for, even though the search engine is constantly gathering information about the search terms that are being used by a registered user of the search engine who is not conducting the searches under the protection of digital anonymity by using special tools that allow for that anonymity. The very fact that someone may use such anonymizing tools demonstrates a reduced alienation and greater mindfulness. Most users do not do that but rely on another central construct of open-market, democratic system that operate within a stated rule of law often based on seminal documents such as a constitution that is the fountainhead of the laws of the land.

Most countries that are based on constitutional laws attempt to grant its citizens some degree of freedom of speech with sufficient guarantee that what one says would not be the reason of harm being brought to the person. For instance, since the adoption of the First Amendment to the American Constitution the notion of freedom of speech has gone through many iterations through the history of the United States, however, the statements of Supreme Court Judge Thurgood Marshall in 1972 very clearly laid out the principles the concept of freedom of speech as, "To permit the continued building of our politics and culture, and to assure self-fulfillment for each individual, our people are guaranteed the right to express any thought, free from government censorship. The essence of this forbidden censorship is content control. Any restriction on expressive activity because of its content would completely undercut the 'profound national commitment to the principle that debate on public issues should be uninhibited, robust, and wide-open.'" This is fundamentally what many people in the USA, and people in other countries with similar principles, believe to be their right, allowing for the alienation and the luxury to be less attentive. However, the notion of freedom of speech alters in the environment when the speech is recorded pervasively and permanently by entities other than the government. These are private institutions that offer the forum for the speech: such as the ability to propagate a micro-blog, place a comment on digital social network, or obtain entertainment by the use of a digital game. This forum is no longer the legendary "Speaker's Corner" in Hyde Park of London, or the University quadrangles across the USA. On the other hand, the forums of digital speech are owned by private corporations who, within legal boundaries, can utilize the speech of individuals in a manner that will be profitable for the institution that offers the forum. This is a new institution, albeit non-governmental, that is a player in the way most people speak today, and quite different from the environment within which Justice Marshall spoke in 1972.

The data in this study, and in other such examples available endlessly in popular media, suggests that the users who are annoyed about the need for attentiveness are still operating within the 1970s paradigm of freedom of speech with the ensuing tension with the need for mindfulness. The resolution of this tension lies in greater mindfulness about the way in which one expresses oneself, especially about oneself, within the digital forums that are owned by private corporations that has a greater interest in advancing the institution's mission, and profit, as opposed to the original principles that motivated the

need to protect the freedom of speech. This recommendation is the point of entry into the second key recommendation related to the way in which privacy of information needs to be reconsidered.

The rapid digitization of information and the willingness with which alienated users of tools such as Facebook have volunteered personal information to the institutions calls into question the way the alienated user views the overall notion of digital privacy. In spite of the many different ways in which privacy has been considered through the ages, there is some agreement that privacy relates to protection. Privacy suggests that people are interested in protecting some aspects of their lives from the public domain and restrict the access of information to the domains specified by the user. Several American laws codify this notion of protection, with the best known one being the HIPAA Privacy Rule enacted by the United States Department of Health and Human Services in 1996 with the aim of protecting the health information that is held by corporate health service providers. Most Americans rely on HIPAA to protect the health information, and therefore remain unconcerned about the privacy of that data, which too has been copiously digitized. As such, there are domains, such as health, where there is a well enunciated statement and assumption of privacy. This, however, is still not true for other domains such digital social networks. It is true that the level of “intimacy” of the information may not be as sensitive as in the case of health data, but the user needs to recognize that the data is not as well protected as in the case of the health data or data that is protected by other laws such the Family Educational Rights and Privacy Act (FERPA) in the case of educational records. It is important to recognize that the presumptions of privacy cannot be applied universally to all digital data.

The findings from this study leads to a recommendation where the user needs to be aware of what data is protected and to what extent so that the user is able to anticipate what may happen if that data were to be exposed to the public domain. This attentiveness to the variability in the privacy of the data is crucial to deciding what a user volunteers to do in terms of sharing data on digital social networks.

FUTURE RESEARCH DIRECTIONS

The challenges with data breaches and the threat to privacy continue unabated as there is growing evidence that the “trusted” systems are not as well protected as users imagine them to be⁶. In addition to the situation examined in this essay, just with Facebook, there have been ongoing issues such as the stealth of millions of “tokens” in the period starting from July 2017 where personal information of users could have been compromised because users were getting used to the ubiquity of “always logged on” state of Facebook and affiliated applications. Although the crisis was precipitated by gaps in the Facebook security system, part of the responsibility also lay with the way in which the users were routinely, without much attentiveness, were enjoying the convenience of not having to log out or not having to memorize multiple credentials. It is imperative that there be ongoing exploration on two fronts.

First, there needs to be a better understanding of the security protocols and the terms and conditions imposed by private corporations in relation to the data of private individuals. This research requires the engagement of technologists and engineers as well as legal professionals. However, this avenue of research would also benefit from the contributions of scholars who explore the ethical components of how private corporations choose to utilize data for the benefit of the corporation. There is little knowledge, from ethical and sociological perspectives, about the way in which a fair balance is achieved between the competing demands of corporate profit and the expectations of privacy on the part of those who choose to use the products provided by the private corporations.

The second focus of research needs to examine the technological knowledge, attitudes and expectations of the user of digital products. In particular, such exploration ought to examine how alienated have the users become from the increasingly ubiquitous technologies, and what they feel about the tools and what their expectations are with respect to issues such as the right to speak and the right to privacy. The attitudinal studies need to focus on the ways in which the notions of voluntary data revelation and the feelings about the ways in which users think that data ought to be used.

In combination, the findings from such ongoing research can help to first build better informed regulatory processes as well as to develop messages and training for the user to allow the users to be better informed about the way in which data is gathered, managed and utilized.

CONCLUSION

There is no indication that three processes will slow down in the near future – 1) voluntary disclosure of data to private corporations without much attention to the process or its consequences, 2) increasing use of corporate platforms such as Facebook to make everyday interaction with the digital World more convenient and 3) the temptation by legitimate and illegitimate entities to gain access to the data. These three forces are all displayed at work in the study described in this essay. The methodological approach offers a unique way to understand the users and offers narrative information that allows for a deeper understanding of the motives and feelings of the users. It is ultimately what the users feel about the issues of convenience, freedom of expression, expectations of privacy and the consequent user activism and demands that will shape the way in which the three forces interact with each other and eventually find a provisional balance.

REFERENCES

- Andrews, J. D. (1982). The structuralist study of narrative: its history, use and limits. In P. Hernadi (Ed.), *The horizon of literature* (pp. 99–124). The University of Nebraska Press.
- Beres, D. (2018). *It's Time to Protect Yourself - and Your Friends - from Facebook*. Mashable.
- Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*.
- DeMers, J. (2018). *What Marketers Can Learn From The Facebook Scandal*. Forbes Magazine.
- Fisher, W. R. (1984). Narration as Human Communication Paradigm: The Case of Public Moral Argument. *Communication Monographs*, 51(1), 1–22. doi:10.1080/03637758409390180
- Fisher, W. R. (1985a). The Narrative Paradigm: An Elaboration. *Communication Monographs*, 52(4), 347–367. doi:10.1080/03637758509376117
- Fisher, W. R. (1985b). The Narrative Paradigm: In the Beginning. *Journal of Communication*, 35(4), 74–89. doi:10.1111/j.1460-2466.1985.tb02974.x
- Khaira, R., Sethi, A., & Sathe, G. (2018). *UIDAI's Aadhaar Software Hacked, ID Database Compromised, Experts Confirm*. Huffington Post.

- Kosinski, M., Stillwell, D. J., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*. 10.1073/pnas.1218772110
- Leskin, P. (2018). The 21 scariest data breaches of 2018. *Business Insider*. Available at: <https://www.businessinsider.com/data-hacks-breaches-biggest-of-2018-2018-12>
- MacIntyre, A. (1981). *After virtue: A study in moral theory* (2nd ed.). The University of Notre Dame Press.
- Meredith, S. (2018). *Here's Everything You Need to Know about the Cambridge Analytica Scandal*. CNBC.
- Meyer, R. (2018). The Cambridge Analytica Scandal, in 3 Paragraphs. *The Atlantic*.
- Mitra, A. (1999). Characteristics of the WWW Text: Tracing Discursive Strategies. *Journal of Computer-Mediated Communication*, 5(1), 0. doi:10.1111/j.1083-6101.1999.tb00330.x
- Mitra, A. (2010). *Alien Technology: Coping with Modern Mysteries*. Sage.
- Mitra, A. (2010). Creating a Presence on Social Networks via Narbs. *Global Media Journal*, 9(16).
- Mitra, A. (2011). Using narratives from social network to manage teacher-student interaction. In Education in a technological world: communicating current and emerging research and technological efforts. Formatex Research Center.
- Mitra, A. (2012a). Collective Narrative Expertise and Using the Narbs of Social Media. In T. Takseva (Ed.), *Social Software and the Evolution of User Expertise: Future Trends in Knowledge Creation and Dissemination*. IGI Global.
- Mitra, A. (2012b). Narbs as a Measure and Indicator of Identity Narratives. In *Investigating Cyber Law and Cyber Ethics: Issues, Impacts and Practices*. IGI Global. doi:10.4018/978-1-61350-132-0.ch008
- Mitra, A. (2013). Mapping Narbs. In G. Wise (Ed.), *New Visualities, New Technologies: The New Ecstasy of Communication*. Ashgate Publishing Ltd.
- Mitra, A. (2014b). *Digital DNA: Managing Identity in Social Networking Sites*. Rupa Publications.
- Mitra, A. (2018). *Needs Assessment: A Systematic Approach to Data Collection* (2nd ed.). Sagamore Publishing.
- Netchitailova, E. (2012). Facebook as a Surveillance Tool: From the Perspective of the User. *TripleC*, 10(2), 683–691. doi:10.31269/triplec.v10i2.404
- Pilgun, M., & Gradoselskaya, G. (2015). Political Communication on Face-book: Russian Case||. *Revista Latina de Comunicación Social*, 70, 726-749.
- Ricouer, P. (1977). The model of the text: Meaningful action considered as text. In F. R. Dallmayr & T. A. McCarthy (Eds.), *Understanding and social inquiry* (pp. 316–334). The University of Notre Dame Press.
- Ricouer, P. (1983). The narrative function. In J. B. Thompson (Ed.), *Paul Ricouer, Hermeneutics and the human sciences: Essays on language, action, and interpretation* (pp. 274–296). Cambridge University Press.
- Ricouer, P. (1984). Time and narrative: Vol. 1 (K. McLaughlin & D. Pellaur, Trans.). The University of Chicago Press.

Rosenberg, M. (2018). How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*.

Staff. (2018). Cambridge Analytica and Facebook: The Scandal so Far. News. *Al Jazeera*.

Weiser, M. (1991). The Computer for the 21st Century. *Scientific American Special Issue on Communications, Computers, and Networks*.

Whittaker, Z. (2018). A new data leak hits Aadhaar, India's national ID database. *ZDNet*.

KEY TERMS AND DEFINITIONS

Big Data: An increasingly popular term that refers to the large amount of digital data that is being accumulated about individuals, where the data is voluminous, collected at great velocity and extremely variegated in detail offering a nearly precise identity narrative of the individual.

Data Breach: A process where large amounts of private data, mostly about individuals, becomes illegally available to people who should not have access to the information.

Identity: The defining elements of an individual composed of items such as age, gender, race as well as the attitudes, opinions, and beliefs of a person.

Narbs: Narbs are little bits of narrative which exist as fragments of information that are almost meaningless in isolation. But when connected together those trace elements of the story produce an identity narrative of an individual (see www.narbs.info).

Narrative: This is the technical term for a story. Humans are story-telling beings and stories offer a way to make sense of the World around us as well as get an understanding of the people who surround us.

Narrative Map: The maps are made up of specific nodes and connectors. The nodes represent attitudes, concepts, behaviors and different issues that people are talking about in their narbs. The map connects the nodal narbs to indicate how strongly nodes are related to each other.

Technological Alienation: Refers to the process where people get distanced from the technology that surrounds them and the tools that they use; the process has led to the development of a law of technological alienation: as technology gets more sophisticated the user perceives it as complicated and gets alienated from the technology (Mitra, 2010).

ENDNOTES

¹ The ideas of alienation are best developed in a series of notes that were never formally published and was called the *Economic and Philosophic Manuscripts* of 1844. The idea of alienation was most carefully developed in these manuscripts and was later included in Marx's best-known work called *Capital* first published in 1867 and thereafter published in many different languages.

² The idea of alienation as developed by Freud appears in his best-known work called *Civilization and Its Discontents* that was first published in 1930 in German and thereafter has become a widely published work in many different languages including a version published by Penguin Books.

³ Erich Fromm developed and published on his take on alienation in the 1955 book called *The Sane Society* that was published by Henry Holt and Company.

- ⁴ The idea of alienation was developed by Durkheim in his book called *Suicide* published in French in 1897, and available now in many different languages including an English version published in 2007 by Penguin Classics.
- ⁵ As stated on the Website www.themediawatch.com, narrative maps are produced from analysis of narbs, where the maps are made up of specific nodes and connectors. The nodes represent attitudes, concepts, behaviors and different issues that people are talking about in their narbs. For instance, a common node is “positive opinion” which could be made up of positively affected terms such as “like,” “love,” “good,” etc. Similarly, another node could be “immigration” referring to the issue. The map connects nodes together to indicate how strongly two nodes are related to each other. For instance, a dark line between two nodes would show a strong relationship between the nodes, suggesting people are talking about two things simultaneously, almost “in the same breath!”
- ⁶ The incident described here is merely the “tip of the iceberg” in terms of data breaches. In 2018 alone, Business Insider reported that there were 21 data breaches that included major corporations such as British Airways, Google+, Marriott Hotels, and the nation-wide database system developed in India called the Aadhar card (Leskin, 2018).

Forensic Acquisition Methods for Cloud Computing Environments

Diane Barrett

Bloomsburg University of Pennsylvania, USA

INTRODUCTION

Cloud computing environments add an inherent layer of complication to a digital forensic investigation. The content of this article explores current forensic acquisition processes, how current processes need to be modified for cloud investigations, and what new acquisition methods can help when it is necessary to garner evidence from a cloud computing based environment. The article is based on an expert panel Delphi study. The purpose of the qualitative Delphi study was to develop a robust contingency framework for deciding when to use traditional forensic investigative practices, when to use modified processes, and when it is necessary to develop new forensic investigative processes more appropriate to the cloud computing environment.

The lack of previous research on how current methods of forensic acquisition processes apply to cloud computing environments has created a gap that may hinder the continued growth of the digital forensics field. In order to garner a true picture of the application of forensic acquisition processes in cloud computing environments, querying a panel of subject matter experts (SMEs) provided the best possible information.

A section will be included that provides a recommendation on how to acquire evidence from cloud based environments while maintaining chain of custody. A final section will include recommendations for additional areas of research in the area of investigating cloud computing environments and acquiring cloud computing based evidence.

BACKGROUND

Cloud Computing Environments

Cloud computing is encompassed in the capabilities of almost all existing technologies. The cloud market is growing at a rate of 20 percent to 25 percent a year, and reached a size of \$127 billion dollars in 2018. Approximately 30 percent of worldwide enterprise applications are offered via the cloud (Kathuria, Mann, Khuntia, Saldanha, & Kauffman, 2018). The concept behind cloud computing is a production environment in which resources and software services do not function locally. Instead, the Internet or the internal network of an organization seamlessly connects numerous host machines running on a virtualized platform (Skemp, 2019).

Joe-Wong and Sen (2018) provide a general layered architecture of cloud infrastructures as a basic model by classifying the architecture into three abstract layers using two models: deployment and service, along with a set of characteristics. The layers from the bottom up are infrastructure, platform, and application. The infrastructure layer provides fundamental computing resources such as processing, storage,

DOI: 10.4018/978-1-7998-3479-3.ch033

and networks. The platform layer delivers higher-level services and abstractions for integration of the ability to perform application functions in the environment. The application layer allows the capability for applications as a service (AaaS).

These three layers are further broken down into service models, deployment models, and attributes. The three well-recognized cloud service models are infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). The four cloud deployment models are community, hybrid, public, and private. The attributes consist of measured and on-demand self-service, resource pooling, rapid elasticity, and broad network access. This is the exact layered architecture outlined by National Institute of Standards and Technology (NIST) in the final issuance of the cloud computing definition dated September 2011.

Environmental Variables

By 2015, as more and more businesses began to sign on as users, Amazon, Microsoft and International Business Machines Corporation (IBM) had all created cloud computing services. Since then, the popularity of cloud computing has increased because organizations are attracted by the affordability, accessibility, and flexibility of the services (Skemp, 2019). The ability to produce dynamic business environments by using cloud computing environments allow organizations of all sizes to respond rapidly to market changes and pursue creative resource saving solutions. Cloud computing environments offer unrestricted scalability and lower data-center setup costs by using multitenancy.

The multitenancy and virtualization characteristics of a cloud computing environment present difficult implementation demands in the areas of security and access control (Fatima, & Ahmad, 2019). The unique security and access control challenges presented by the use of multitenancy and virtualization in cloud computing environments exist because many individual environments share the same set of hardware. The sharing of storage blocks can result in the accidental and unauthorized flow of information (Fatima, & Ahmad, 2019). The diversity of services offered in cloud computing environments requires variable levels of granularity when implementing access control mechanisms. The risk of resource exploitation by unauthorized users is significantly increased when there are insufficient or untrustworthy authorization mechanisms implemented in a cloud computing environment (Chaudhary & Siddique, 2017).

Cloud computing environments offer many organizational benefits by providing scalability and elasticity but come with complex computing infrastructures. Every cloud deployment and service model instance is different. For example, one SaaS implementation can be completely different from the next. There are many challenges associated with the use of cloud computing environments and existing issues are only beginning to be addressed. As more and more businesses move to a cloud computing environment, automated service provisioning, virtual machine migration, server consolidation, and security are topics that have garnered the attention of the research community.

Digital Evidence Seizure

Digital forensics focuses on the retrieval and analysis of data found on digital devices relative to some type of unauthorized or criminal activity. Traditional digital forensics processes consist of crime scene evidence collection, evidence preservation, evidence analysis, and presentation of the analysis results (Mansour, 2016).

Current traditional digital acquisition processes include maintaining chain of custody control of forensic evidence data. This chain of custody control occurs in the evidence collection phase through the imaging of a system (Shah, Saleem, & Zulqarnain, 2017).

Cloud computing technology disrupts the initial step of evidence collection in conducting a digital forensic investigation. One of the main drawbacks is data acquisition (Monsour, 2016). Traditionally, the seizure of forensic evidence occurs through a search warrant or other legal request. In a cloud computing environment, the emphasis is on the cloud provider contract as opposed to a search warrant or legal request because the data can be geographically disbursed, is under the control of the service provider, and data are spread out amongst numerous servers and other components (Fatima, & Ahmad 2019). In some cases, the forensic investigator never gets to acquire the evidence. The designated employee at the service provider or the cloud broker performs the evidence acquisition and supplies the information to the client, who then sends the information to the investigator (Pichan, Lazarescu, & Soh, 2015). This is contrary to traditional forensic practice where legal documents provide the case investigator the ability to acquire the evidence.

When a crime happens, the location of recoverable data includes the client machine, the equipment of the Internet Service Provider (ISP), the cloud service provider equipment, and the data backups of the cloud service provider. This makes acquiring evidence from a crime committed in a cloud computing environment much more difficult since the environment is not localized like it is in a traditional digital forensic evidence acquisition (Arford, E. 2017). Traditional imaging cannot be performed because it is not possible to take down and create a forensic image of such a large environment.

The digital forensics discipline faces quite a few challenges where cloud computing is concerned especially during the acquisition phase (Mansour, 2016). Traditional forensic evidence acquisition processes do not fit well into cloud computing because of the way cloud computing works. Traditional digital forensic acquisition processes focus on individual computers and isolated environments, while cloud computing forensic acquisition processes include the intricacies of complex infrastructures including virtual servers, applications, and diverse operating platforms that may be located in foreign countries (Burney, Asif, & Abbas, 2016).

Common forensic procedure is to shut down the system in order to take a forensic image of the system. This procedure cannot apply to hosted cloud services due to disruption of service to a wide scope of users. Cloud computing systems consist of multiple user environments, using a variety of services. Shutting down a cloud computing system disrupts services to all the user environments hosted on the system.

When the user or service provider shuts down the virtual machine because of inactivity, service providers reallocate the storage space originally allocated to the virtual machine to other virtual machines (Chaudhary & Siddique, 2017). Upon deletion of the virtual machine and the overwriting of storage space, the evidence disappears. Although it is possible to recover forensic data from a running virtual machine, if the files associated with the virtual machine are not collected, valuable evidence is missed. Associated virtual machine files contain information such as configurations settings and memory content are often not collected as part of the acquisition process.

One of the most commonly used commercial cloud services is Amazon Elastic Compute Cloud (EC2). Hosting cloud services on geographically diverse servers using distributed file systems provides scalability and reliability but present forensics challenges (Alqahtany, S., Clarke, N., Furnell, S., & Reich, C. (2015). Forensic analysis of distributed file systems is difficult (Teing, Dehghantanha, & Choo, 2017). Multi-national web-based organizations such as Facebook and Yahoo use the Hadoop file system. Hadoop is a common distributed file system used in large data environments such as cloud computing. Due to the way the Hadoop file system works, it processes petabytes of data and erasure of forensic evidence could easily go without notice (Teing et al., 2017). The development of Hadoop occurred with fast processing of data in mind and there was very little consideration given to security or forensic analysis capabilities.

CLOUD ACQUISITION PROCESSES

Without a clear understanding of how cloud computing works and the underlying infrastructure components, law enforcement agencies cannot properly investigate cloud computing environments. Some forensic practitioners have taken the view that the acquisition of evidence from the cloud is merely an extension of existing actions used to acquire data from a network because the cloud data is stored similar to network data (Adams, 2013). Contrary to this view, there are differences between cloud computing forensics and network forensics. The primary similarities and differences between cloud computing forensics and network forensics stem from the key areas of physical access, logical access, segregation of evidence, methodology, and legal considerations.

There was no prior research conducted on this subject. In order to garner practitioner opinions on cloud forensics and the application of traditional forensic acquisition methods to cloud forensic environments, an expert panel Delphi study was conducted. The Delphi method provides a means to refine a group consensus opinion from expert practitioners using a sequence of question rounds. Expert panel member selection was based on the criteria from a submitted statement of qualifications. Only digital forensic investigators with at least five years of relevant field experience, published work, industry presentations, and recognition were eligible to participate in this study. The expert panel consisted of 14 members from several countries.

An online written narrative interview questionnaire for the study began with 10 open-ended written questions on cloud computing based on the cloud study by Ruan Baggili, Carthy, & Kechadi (2011). Panel members were then asked to evaluate 20 common forensic procedures for applicability to cloud computing environments. The common forensic procedures selected are shown in Figure 1.

The findings demonstrated there were very diverse opinions on cloud computing, cloud forensics, and the effect cloud computing environments had on digital forensics. Standard evidence acquisition procedures, federal and local laws, court accepted methods, and the cooperation of the cloud provider were all factors that affected the way a successful forensic acquisition was conducted in a cloud computing environment. The areas of tools, processes, and guidance available for forensic evidence acquisitions in cloud computing were relatively immature.

A recap of the responses to the evaluation of the 20 common forensic procedures for applicability to cloud computing environments indicated several key points. Only 55% of the 20 pre-selected traditional forensic processes were usable for the forensic acquisition of digital evidence in cloud computing environments and the usability of those processes had some limitations. Post-acquisition processes were most suited for application in cloud computing environments. Following post-acquisition processes in order of applicability were live acquisition processes. The main limitation in this area was access to the cloud server in order to perform the processes. The results analysis suggested this category of processes had solutions for digital evidence acquisitions in cloud computing environments because the processes were modeled after already established network forensic processes. Thirty-five percent of the 20 pre-selected traditional forensic processes were modifiable for the forensic acquisition of digital evidence in cloud computing environments depending on the level of access and service provider cooperation. Pre-acquisition processes were most suited for modification in cloud computing environments. Five percent of the 20 pre-selected traditional forensic processes required the development of new processes for the forensic acquisition of digital evidence in cloud computing environments. The panel members suggested that pursuing the development of new processes in some cases was moot because the processes were irrelevant to cloud computing environments.

The panel experts agreed that current acquisition processes in the category of dead drive acquisitions did not fit into cloud computing environments and many of the processes should no longer be included because they were not applicable or modifiable.

Fifty-seven percent of the panel members identified remote access as a theme; while many of the remaining members argued that cloud computing does not always involve remote access as in a hybrid cloud. The panel members identified three main characteristics of defining cloud forensics. Verification of data integrity, working with service providers to obtain required information and legal admissibility, and international cooperation were the three main characteristics.

CLOUD ACQUISITION APPROACHES

Ruan, Carthy, and Kechadi (2012) proposed that cloud forensics is a subsection of network forensics and a cross-discipline between cloud computing and digital forensics, adopting the same definitions of cloud computing and digital forensics provided by NIST. The methodology used originates at three distinct dimensions: legal, technical, and organizational. Choosing this methodology encompasses jurisdictional issues, the lack of international cooperation, and formation of foundational standards and sound policies. Ruan et al. (2012) developed a high-level conceptual model that analyzed the forensic impact of cloud computing environments based on deployment models, service offerings, cloud actor roles, segregation of duties, and forensic artifacts obtainable from the environment. Using log files, network monitoring and investigation of malicious behavior can occur (Tak, Kachhwaha, & Mahia, 2018). This area of cloud forensics has the most research and approaches since log files are readily available in most cases.

Adams (2013) proposed the advanced data acquisition model (ADAM). The model breaks down forensic processes into three main functions derived from analysis of previous models. The actions in the first stage occur upon notification of the scope of work to the investigator. The activities in this stage happen prior to arriving at the incident location. In the next stage, the investigator prepares for the data acquisition upon entry of the incident scene. The final stage consists of acquiring and handling the evidentiary data.

Adams (2013) used a distinctions, systems, relationships, and perspectives (DSRP) thinking patterns in creating ADAM, tying scientific thinking to the forensic processes. Initial testing of ADAM compared ADAM to the processes used in three previous investigations. The results produced a viable model. Cases in Australia and New Zealand are using process flowcharts as a manner in which to help the jury and court understand scientific forensic evidence. Widiyasono, Riadi, and Luthfi, (2016) conducted and documented an investigation on the services of private cloud computing using the ADAM method. The methodology applied to a private cloud environment worked well and is a viable methodology for this type of environment.

Shende (2010) approached the digital investigations of cloud computing environments as the mapping of computational and storage structures falling within the investigative scope. Emphasizing the cross boundaries of responsibility and access, service level agreements play a large part in the ability to investigate a cloud computing environment. The key to data access in cloud computing investigations lies in the dialogue between the organization and cloud service vendor in establishing incident response requirements, processes, and protocols when drafting a service level agreement (SLA) (Shende, 2010).

Lu, Lin, Liang, and Shen (2010) proposed using a secure source structure based on the bilinear pairing technique to deliver reliable digital forensic evidence in cloud computing environments. Using a cloud computing architecture comprised of a cloud service provider (SP), a trusted system manager (SM),

and an enterprise size user environment, the secure provenance SP method demonstrates confidentiality of information, authorized access, anonymous authentication, and ownership tracking. Extending this method for forensic use provides ownership information associated with every version of a document to the SM. The SM then uses an ownership-tracking algorithm to follow the evidence to a unique user identity. This proposed design encompasses a wide range of security features, making it ideal for use in the forensic investigation of cloud computing environments.

Wittek and Daranyi (2012) took an approach similar to Lu et al. (2010), in that the work focused on the tracking of digital assets. Using a digital preservation (DP) policy methodology ensures that digital objects remain valid over an extended time period. Cloud computing infrastructure provides a platform ad-hoc computation along with distributed digital preservation. Wittek and Daranyi (2012) proposed a middleware solution in the business services layer of the Sustaining Heritage Access through Multivalent Archiving (SHAMAN) core infrastructure. The business layer serves as the intermediary between the top presentation layer and the lower infrastructure layer. The proposed solution has promising future implications for digital preservation.

Haque, S., Atkison, T. (2018) devised a forensic enabled data provenance model for public cloud in which a cloud auditor is added to the environment and basically performs analysis on the cloud services and provides feedback. This methodology has already been implemented in a more basic model for monitoring cloud elements.

More recently, researchers have been concentrating on separate areas for cloud evidence. Burney, Abbas, & Mumahhad, (2016) took an approach that divides the evidence acquisition process by the cloud environment. For example, a computer forensics investigator can get instance access to log history of the compromised system in SaaS and PaaS service models whereas in IaaS model an image of Virtual Machine can be taken for investigation.

Cooperative storage cloud forensic is relatively new and is just beginning to be explored. Teing, Dehghantanha, Choo, Dargahi, & Conti, (2017) have formulated processes for cooperative storage based on client side analysis while Ho, Kao, and Wu (2018) approach cloud forensics by using timestamp pattern identification.

MAINTAINING CHAIN OF CUSTODY

The most important element in the digital forensics is to maintain the integrity and the chain of custody of the digital evidence. A break in the chain of custody simply means that the case is lost in a court of law (Simou, Kalloniatis, Gritzalis, & Mouratidis, 2016).

Cloud environments are difficult to access in a forensic manner because its environment is live and the evidence cannot be logged into directly as it violates preserving the state of data and alters the data state. This can be compared to looking through a hard drive to find evidence without first creating a forensic image of the drive. The first rule of evidence is to never work on original evidence.

A cloud based investigation may include many cloud components including cloud based email, cloud based personal storage, social media, and cloud based corporate storage. The evidence items can encompass personal e-mail accounts, Facebook pages, corporate storage buckets on Amazon Web Services (AWS), and personal storage on DropBox, Box, and Google Drive. Privacy and legal considerations are a main concern since some of the storage buckets may be located in foreign countries.

Using an investigative environment consisting of a virtual machine (VM) that contains several forensic tools such as Access Data's FTK imager, F-Response Universal, and Paraben's E3 DS, evidence

was able to be acquired while maintaining chain of custody. Tools such as Paraben E3 and F-Response access the cloud environment through an authentication API. Basically the tools act as an intermediary between the forensic examiner and the cloud environment. This allows evidence to be mounted as read-only and prevents direct access by the examiner. Then standard validation processes such as hashing can be conducted. This is an acceptable process from a forensic standpoint. The process does not allow the investigator to touch the original evidence, reinforcing proper forensic procedures. An API-centric approach to tool development yields forensically sound results that far exceed the type of information available by client-side

Analysis (Roussev, Ahmed, Barreto, McCulley, & Shanmughan, 2016).

Once the environment is accessed, both F-Response Universal, and Paraben's E3 DS can compress, hash, and export the data. The evidence is preserved in a forensic manner using this process. Once the acquisition is complete, the evidence is analyzed in the same manner as any other evidence. This validates the expert panel findings that post-acquisition processes were most suited for application in cloud computing environments. Cloud acquired evidence is analyzed in the same manner as any other acquisition. This process is supported by prior research from the 14-member expert panel survey. There are a few important points worth mentioning. Credentials are needed in order to access the environment. The authentication of the accounts require logins and passwords. Authentication keys are required to access AWS storage. As long as all parties are cooperative this information will be available. If the parties are not cooperative, the process cannot be used. When parties are not cooperative, any investigation is impeded, whether it is a cloud-based or a traditional forensic based investigation. When two-factor authentication is used, the process will be difficult as currently tools are not set-up to access information when two-factor authentication is required. This difficulty is encountered whether it is a cloud-based or a traditional forensic based investigation. Privacy and legal issues are a consideration, especially since the passage of General Data Protection Regulation (GDPR).

SOLUTIONS

The fundamental issue in this study was the effect on information technology digital forensic acquisition processes by organizational adoption of cloud computing technology. The study methodology used to develop a new forensic choice set consisting of information-contingent plans for choosing actions based on the uncertainty of information was conceptual and based on existing frameworks. The resulting framework offers a realistic view of cloud computing environment variables that allows forensic acquisition process decisions to be made without making prior assumptions that traditional forensic processes must apply. The contingencies examined are the NIST defined cloud computing service and deployment models applied to current forensic acquisition processes.

The basis for the underlying theoretical application associated with the research study relied on contingency theory to provide a framework for determining when to use traditional forensics acquisition processes, when to use modified processes, and when the development of new methods is required for forensic evidence acquisitions in cloud computing environments. The applicable theoretical framework used in linking this study to other information security research using contingency theory was the framework proposed by Austin and Devin (2009). The research extends the analytical framework of Austin and Devin (2009) by creating a framework for cloud computing forensic acquisition processes based on the contingency framework for determining when to use plan-based methods and when to use agile methods.

The main agenda in doing this research was bringing a framework based on subject matter expert opinions to the attention of forensic examiners on the applicability of current forensics evidence acquisition procedures to cloud computing environments. Since there is a lack of knowledge and understanding about the applicability of forensic evidence acquisition processes in cloud computing environments by examiners, the goal was to provide enlightenment in this area, making forensics investigations more productive and the prosecution of criminal activity more likely.

FUTURE RESEARCH DIRECTIONS

The opportunity exists for researchers to make innovative contributions and have a substantial impact on the cloud computing industry. One of the significant contributions of the panel study is the identification of contingency factors such as available tools, access, availability, and acquisition scope as the underlying elements when choosing the application of digital forensic methods to cloud computing environments. These contingency factors are easily ported to other evidence acquisition methods for expanding research in cloud forensics since there are such diverse environments. Research can be conducted on the development of new processes where multitenancy and physical access are concerned. Setting new standards and best practices when these conditions exist will be difficult for digital forensic practitioners because other parties must be relied upon to provide evidence. There are ample opportunities available to study specific identified deployment models, service offerings, and obtainable forensic artifact areas.

CONCLUSION

The article is based on an expert panel Delphi study. The purpose of the qualitative Delphi study was to develop a robust contingency framework for deciding when to use traditional forensic acquisition practices, when to use modified processes, and when it is necessary to develop new forensic acquisition processes through the evaluation of 20 conventionally recognized forensic acquisition processes by a panel of SMEs.

The study findings indicate that cloud investigations encompass many evidence items. As an industry, digital forensics is beginning to formulate proper acquisition processes of digital evidence in cloud computing environments. Many of the digital acquisition processes that apply to cloud computing environments are modeled after already established network forensic processes. The privacy and legal considerations associated with cloud-based evidence must be addressed. In setting standards for proper acquisition of forensic evidence in cloud computing environments, organizations must create legal requirement documentation encompassing both vendor contracts and the organization. After the legal requirements are set, digital forensic practitioners have the availability to set best practices, but those best practices will be contingent upon the environment and existing conditions. Tools for cloud-based evidence acquisitions that are different from traditional digital forensic tools are now being developed.

REFERENCES

- Adams, R. (2013). The emergence of cloud storage and the need for a new digital forensic process model. In K. Ruan (Ed.), *Cybercrime and cloud forensics: Applications for investigation processes* (pp. 79–104). doi:10.4018/978-1-4666-2662-1.ch004
- Alqahtany, S., Clarke, N., Furnell, S., & Reich, C. (2015). Cloud forensics: A review of challenges, solutions and open problems. *Cloud Computing (iccc)*, 2015 international conference, 1–9.
- Arford, E. (2017). *Holes in the Cloud Weaknesses in Applying Digital Forensic Processes to Cloud Computing*. *Computer Security*. Tufts University.
- Austin, R. D., & Devin, L. (2009). Weighing the benefits and costs of flexibility in making software: Toward a contingency theory of the determinants of development process design. *Information Systems Research*, 20(3), 462–479. doi:10.1287/isre.1090.0242
- Burney, A., Asif, M., & Abbas, Z. (2016). Forensics Issues in Cloud Computing. *Journal of Computer and Communications*, 4(10), 63–69. doi:10.4236/jcc.2016.410007
- Chaudhary, O., & Siddique, A. S. (2017). Cloud Computing Application: Its Security Issues and Challenges Faced During Cloud Forensics and Investigation. *International Journal of Advanced Research in Computer Science*, 8(2), 12–15.
- Fatima, S., & Ahmad, S. (2019). An Exhaustive Review on Security Issues. *Cloud Computing. Transactions on Internet and Information Systems (Seoul)*, 13(6), 3219–3237. doi:10.3837/tiis.2019.06.025
- Haque, S., & Atkison, T. (2018). A Forensic Enabled Data Provenance Model For Public Cloud. *Journal of Digital Forensics. Security & Law*, 13(3), 47–65.
- Ho, S. M., Kao, D., & Wu, W.-Y. (2018). Following the breadcrumbs: Timestamp pattern identification for cloud forensics. *Digital Investigation*, 24, 79–94. doi:10.1016/j.diin.2017.12.001
- Joe-Wong, C., & Sen, S. (2018). Harnessing the Power of the Cloud: Revenue, Fairness, and Cloud Neutrality. *Journal of Management Information Systems*, 35(3), 813–836. . doi:10.1080/07421222.2018.1481639
- Kathuria, A., Mann, A., Khuntia, J., Saldanha, T. J. V., & Kauffman, R. J. (2018). A Strategic Value Appropriation Path for Cloud Computing. *Journal of Management Information Systems*, 35(3), 740–775. doi:10.1080/07421222.2018.1481635
- Lu, R., Lin, X., Liang, X., & Shen, X. S. (2010, April). Secure provenance: the essential of bread and butter of data forensics in cloud computing. *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, 282–292. 10.1145/1755688.1755723
- Mansour, S. (2016). Cloud Computing and Digital Forensics Challenges. *GSTF Journal on Computing*, 5(1), 116–120. doi:10.5176/2251-3043_4.4.361
- Pichan, A., Lazarescu, M., & Soh, S. (2015). Cloud forensics: Technical challenges, solutions and comparative analysis. *Digital Investigation*, 13, 38–57. doi:10.1016/j.diin.2015.03.002
- Roussev, V., Ahmed, I., Barreto, A., McCulley, S., & Shanmughan, V. (2016). Cloud forensics–Tool development studies & future outlook. *Digital Investigation*, 18, 79–95. doi:10.1016/j.diin.2016.05.001

Ruan, K., Baggili, I., Carthy, J., & Kechadi, T. (2011, May). Survey on cloud forensics and critical criteria for cloud forensic capability. *Journal of Digital Forensics, Security and Law, Conference Proceedings*, 55-70.

Ruan, K., Carthy, J., & Kechadi, T. (2012, May). NIST cloud computing reference model and its forensic implications. *Journal of Digital Forensics, Security and Law, Conference Proceedings*. Retrieved from http://www.digitalforensics-conference.org/subscriptions/proceedings_2012.htm

Shah, M. S. M. B., Saleem, S., & Zulqarnain, R. (2017). Protecting Digital Evidence Integrity and Preserving Chain of Custody, *Journal of Digital Forensics, Security and Law*, 12(2), 12. Advance online publication. doi:10.15394/jdfsl.2017.1478

Shende, J. (2010). Live forensics and the cloud. *DFI News*. Retrieved from <http://www.dfinews.com/article/live-forensics-and-cloud-part-1>

Simou, S., Kalloniatis, C., Gritzalis, S., & Mouratidis, H. (2016). A survey on cloud forensics challenges and solutions. *Security and Communication Networks*, 9(18), 6285–6314. doi:10.1002/ec.1688

Skemp, K. M. (2019). *Cloud computing*. Salem Press Encyclopedia of Science.

Tak, V., Kachhwaha, R., & Mahia, R. N. (2018). Secure Log Forensics as a Service in Cloud Computing. *International Journal of Advanced Research in Computer Science*, 9(1), 544–546. doi:10.26483/ijarcs.v9i1.5373

Teing, Y.-Y., Dehghantanha, A., & Choo, K.-K. R. (2017). CloudMe forensics: A case of big data forensic investigation. *Concurrency and Computation*, 20(5). Advance online publication. doi:10.1002/cpe.4277

Teing, Y.-Y., Dehghantanha, A., Choo, K.-K. R., Dargahi, T., & Conti, M. (2017). Forensic Investigation of Cooperative Storage Cloud Service: Symform as a Case Study. *Journal of Forensic Sciences*, 62(3), 641–654. doi:10.1111/1556-4029.13271 PMID:27885663

Widiyasono, N., Riadi, I., & Luthfi, A. (2016). Investigation on the Services of Private Cloud Computing by Using ADAM Method. *Iranian Journal of Electrical and Computer Engineering*, 6(5), 2387–2395.

Wittek, P., & Darányi, S. (2012). Digital preservation in grids and clouds: A middleware approach. *Journal of Grid Computing*, 10(1), 133–149. doi:10.1007/10723-012-9206-7

Zhou, X., & Mao, F. (2012, August). A semantics web service composition approach based on cloud computing. *Fourth International Conference on Computational and Information Sciences (ICCIS)*, 2012, 807-810. doi:10.1109/ICCIS.2012.43

ADDITIONAL READING

Burney, A., Asif, M., & Abbas, Z. (2016). Forensics Issues in Cloud Computing. *Journal of Computer and Communications*, 4(10), 63–69. doi:10.4236/jcc.2016.410007

Chaudhary, O., & Siddique, A. S. (2017). Cloud Computing Application: Its Security Issues and Challenges Faced During Cloud Forensics and Investigation. *International Journal of Advanced Research in Computer Science*, 8(2), 12–15.

Fatima, S., & Ahmad, S. (2019). An Exhaustive Review on Security Issues in Cloud Computing. *Transactions on Internet and Information Systems (Seoul)*, 13(6), 3219–3237. doi:10.3837/tiis.2019.06.025

Liu, C., Singhal, A., & Wijesekera, D. (2016). A probabilistic network forensic model for evidence analysis. In G. Peterson & S. Shenoi (Eds.), *Advances in Digital Forensics XII* (pp. 189–210). Springer. doi:10.1007/978-3-319-46279-0_10

Liu, C., Singhal, A., & Wijesekera, D. (2017). Identifying evidence for cloud forensic analysis. In G. Peterson & S. Shenoi (Eds.), *Advances in Digital Forensics XIII* (pp. 111–130). Springer. doi:10.1007/978-3-319-67208-3_7

Roussev, V., Ahmed, I., Barreto, A., McCulley, S., & Shanmughan, V. (2016). Cloud forensics—Tool development studies & future outlook. *Digital Investigation*, 18, 79–95. doi:10.1016/j.diin.2016.05.001

Ryder, S., & Le-Khac, N. (2016). The end of effective law enforcement in the cloud? To encrypt or not to encrypt, Proceedings of the Ninth IEEE International Conference on Cloud Computing, pp. 904–907

Shah, M. S. M. B., Saleem, S., & Zulqarnain, R. (2017). Protecting Digital Evidence Integrity and Preserving Chain of Custody. *Journal of Digital Forensics, Security and Law*, 12(2), 12. Advance online publication. doi:10.15394/jdfsl.2017.1478

KEY TERMS AND DEFINITIONS

Cloud Computing: Cloud computing is an environment in which the Internet or the internal network of an organization seamlessly connects numerous host machines running on a virtualized platform to provide resources and software services.

Digital Forensics: Digital forensics is the division of forensic science that focuses on investigating and analyzing information artifacts located on digital devices involved in computer crime cases.

Elasticity: Elasticity is the dynamic allocation of cloud resources so that an organization can use the exact amount of resources required during any given time period.

Infrastructure as a Service (IaaS): IaaS is an Internet delivered virtual environment in which the service provider supplies cloud infrastructure services that can include software, hardware, and networking equipment.

Multitenancy: Multitenancy is a software architecture in which one server hosts a multitude of different virtual environments for unrelated organizations, or tenants.

Platform as a Service (PaaS): PaaS is an Internet delivered virtual environment in which the service provider supplies cloud platform services that include the operating system and other platform-associated services.

Service-Level Agreement (SLA): An SLA is a binding contract between service providers and purchasing organizations that specifies the service parameters and recourse if an unexpected business interruption occurs.

Software as a Service (SaaS): SaaS is an Internet delivered virtual environment in which the service provider supplies commercial applications, eliminating associated purchasing and installation costs for the organization.

Virtual or Virtualized: Virtual or virtualized is a term used to explain the process of running an environment other than the native computing environment.

Threats and Vulnerabilities of Mobile Applications

Thangavel M.

 <https://orcid.org/0000-0002-2510-8857>

Thiagarajar College of Engineering, India

Divyaprabha M.

Thiagarajar College of Engineering, India

Abinaya C.

Thiagarajar College of Engineering, India

INTRODUCTION

Today's mobile smartphones are very powerful, and many mobile applications use wireless multimedia communication. The attack faced by the mobile phones are increasing day by day. The downloadable mobile application has many types of security issues. Malicious apps look like fine in downloadable sites, but they specially designed to commit an attack. Application-based threats may generally be fit into the following categories: Malware, Spyware, Physical Threats, and Vulnerable Application. Malware is a software that performs a malicious action when it is installed without your knowledge. It can even make changes to your phone bill and also an attacker can gain control over it. Spyware is mainly designed to collect all the private information of the system without your knowledge. It specifically targets the browsing history, contact list, user location, email, etc. Privacy threats are somewhat similar to Spyware, but it gathers sensitive information rather than performing their action. Vulnerable applications allow an attacker to access sensitive information, perform undesirable actions, stop service from functioning correctly, or download apps to your device without your knowledge. As mobile devices are constantly connected to the internet and frequently used to access web-based services, web-based threats will cause persistent issues for mobile devices. Then server-side attacks to mobile devices are also increasing, any communication that happens between the application and the user outside the mobile phones happens through a server. Thus, this becomes a primary target that gets exploited by hackers. Some common reasons for this are small security budgets, lack of security knowledge in a new language, too much dependability on the mobile OS for security updates and responsibility, and vulnerabilities due to cross-platform development and compilation. The easiest way to overcome this server vulnerability is to scan mobile devices. Other vulnerable components of the mobile devices are incorrect permission setting that allows access to controlled functions such as the camera or GPS, and obfuscation where functionality or processing capabilities are hidden or obscured from the used emails, email addresses, and attached data and malformed SMS causing devices to crash. Sometimes untested mobile applications also cause the attack to the device, that is the users tend to download apps from third-party vendors instead of downloading from the app store; many of these apps don't have a clear source and becomes vulnerable due to the coding errors, which are not updated enough to prevent exploitation. The excessive power consumption applications are also vulnerable to the mobile devices that run continuously in the background, that drain the battery thereby reducing the system availability.

DOI: 10.4018/978-1-7998-3479-3.ch034

BACKGROUND

Based on the research, the chapter shows that the major attacks are phishing attacks, attacks based on the use of phone cameras, and a screenshot attack “capture me”. By this threat of the mobile application, the user credentials are stolen by an attacker. Several recent works have studied the issue of obtaining private information on smartphones using multimedia devices such as microphones and cameras if the camera is affected by malicious software. Moreover, most of the users have the impression that smartphones are just the phones installed with a variety of software and don’t think of them as handheld computers that are more vulnerable to cyber-attacks. A recent survey reported that 267,259 malware-infected apps have been found, among which 254,158 reside on the Android platform (Daojing, 2015). It also suggested that the number of malware in apps has increased by 614% since 2012. The wide use of smart devices (phones and tablets) encourage the financial institution to consider mobile banking applications as a necessary service to their clients. This issue should be addressed by both preventive approaches and effective detection techniques. For preventing, the application developer should ensure this app by policies governing secure coding and privacy, which do not access any unnecessary information. Then the App market administrator should strictly verify every uploaded application and remove the suspicious app. Detecting techniques are based on signature-based or anomaly-based. In signature-based, the malware is identified by comparing the signatures with the known signature, and in anomaly-based, the normal behavior of the system is monitored and it checks for deviations from the normal behavior.

THREATS

In information security, a threat is a danger or a possibility for an attack that might exploit a vulnerability to breach the security system and therefore cause huge damage to the system. It can be either intentional or accidental. As the population of smartphone users is rapidly increased, the hackers are targeting mobile phones, which have a wide variety of applications. There are many types of threats that are classified as application-based, web-based, and server-side threats (Delac, 2017).

APPLICATION-BASED THREATS

The number of malware, spyware has been increased in mobile devices by more than 600% since 2012, with over 250,000 malware in the Android operating system alone (Daojing 2015). The malware is malicious software that refers to viruses, worms, Trojan horses, ransomware, which cause the deviation from the normal behavior of the system. The malware has many types of malicious behaviors which include leaking sensitive data or stealing credentials, restricting device usage, mobile distributed denial of service, etc. The recent malware is the root exploits (jailbreak exploits) by which the attacker gain access privileges. In mobile, the official application stores such as Google Play and Apple App Store will scan the application they offer for traces of malware, spyware, and remove the application if any threat is detected. The main threat lies in application installation and application handling in mobile devices. The security of the mobile users won’t end once an application has been installed, then the user is asked to make a decision regarding the security and privacy while handling the application. Moreover, it will ask the user for permission to access the information in the mobile device. The application installed are tends to update automatically, in that time it can even change the permission settings, which become a major threat.

Threats in the Android Platform

The devices which use the Android operating system are popular with the customers. Malware attack is the biggest problem we are facing in the android platform. The risk of threat has become so high, the devices that access the internet are subjected to attack or infection by Trojans, keyloggers, and other forms of malware and spyware.

Threats in iOS Platform

All Apple users think that their systems are provided with high security, but the reality is iPhone and other Apple mobile devices also become a target for the attacks. In September 2015, one of the most serious threats occurred, when it was discovered a large number of trojanized iOS apps were being hosted on the service (Sheik, 2020). Here the apps were developed using a malicious version of Xcode. In addition to this, the organized apps are capable of receiving commands from C&C servers to carry out phishing attacks. One of the major threats in iOS is jailbreaking, by which the attacker can escalate the privileges by removing the software restrictions imposed by Apple on iOS. It can be done using kernel patches.

Threats in Windows Platform

The mobile devices with Windows OS are used less in numbers when compared with other Android and iOS, due to its compatibility. It also subjected to many types of attacks mainly the command injection attacks thereby stealing the credentials. currently, Microsoft is seeking to unify their applications, which means that many developers are remaking their applications as universal Windows apps as the ROI on maintaining a stand-alone windows phone app is too little. The Windows App Store will have only fewer applications than in android.

WEB-BASED THREATS

In today's world, the mobile device offers a high level of convenience. At any location (office, home, hotel, playground, road, or any place in the world) the mobile user can fulfill their daily needs, including buying, communication searching, selling, entertainment, making payments, etc. through web applications in mobiles. Threats related to such activity is a major concern and it became a serious problem for the mobile users. This extreme level of comfort brought an extreme level of security issues and threats. Some of the web-based threats are described below (Seo, 2011).

Mobile Browsing

The best feature of the mobile device is browsing, which provides the best use of internet applications. Normally in mobiles, the user cannot see the entire web address or the domain address; hence, it is difficult to check whether the URL is secure or not. It becomes a threat to a phishing-related attack.

Multiple User Logging

Due to the progressive growth of social networking sites and single sign-on in the mobile application ecosystem, it has been found that 60% of the applications are insecure due to the use same login credentials for multiple website or applications moreover in mobile devices the user login to the application once, and they never log out mainly in Gmail account, which becomes a serious threat to mobile devices (Khan, 2015) by which the attackers can acquire the credentials and can gain access to a user's profile page.

Client-Side Injection

As mobile devices are accessing the internet for every action there may be a chance of executing the malicious program without the knowledge of the user, it occurs by mobile applications or web browsing client-side injection. There are many types of client-side injection some of them are HTML injection, SQL injection, etc. In this way, any data can be injected.

Improper Session Handling

Session handling is identified as an important security concern for web applications in mobile devices. Improper session handling is a major threat when mobile devices are accessing the internet over any platform. the application with long expiry sessions is often subjected to the attack, especially in a mobile device. This poor management session provides a way for an attacker to unauthorized access by session hijacking

Weak Authentication and Brute Force Attack

Today, almost every application relies on password-based authentication. The who are using this application won't enforce strong passwords and do not provide any security for the application. Thus, the user itself exposes themselves to a host of threats for an attacker. With the brute force attack, the hacker gains the credentials of the user due to the weak authentication mechanism.

Web-Based Threats in Different Mobile Platforms

The possibility of attacks is larger in the mobile devices of Android platforms when compared with other operating systems. In Apple iOS only a few applications are free of cost, and other applications can be used only by paying money; hence, the user will be considered that there can be some security feature in it. Then the Windows and Blackberry OS are rare in usage by the user, but in every platform, the possibility of an attack is increasing day by day.

NETWORK-BASED THREATS

The mobile devices provide the best support to cellular networks and wireless LAN IEEE 802.11. In both the networks, we have different types of threats for mobile users. Some of the network-based threats are described as follows (Razzaq 2013):

Network Exploits

This threat occurs in mobile devices due to the faults in the operating system and other applications in the mobile. When the mobile device is connected to the network, there is a possibility of installing some malicious software on the user's device without the knowledge of the users.

Mobile Network Services

The network provides a lot of services like SMS, MMS, voice call, multimedia transmission, etc. Using these services, the hacker can launch the attack. For example, phishing attack can be launched by the attacker by which he impersonates himself as a legitimate user and communicates with the user.

Wi-Fi Sniffing

It is one of the major threats to wireless technology. Here every device communicating in the wireless medium is a target for the attack. By intercepting the communication (i.e., man in the middle attack) or by simply monitoring the attacker gain access. Then by placing a rogue access point with a strong signal, the hacker can steal sensitive information. Therefore, communication in the wireless medium is a big risk, and unencrypted data can easily be stolen by the attackers.

ATTACKS

The attack on the mobile device occurs due to the vulnerabilities and threats that we have seen before. The attack can occur in three areas they are, in the mobile device, in the network, then finally in the database (Xenakis, 2015). The attacks are broadly classified into active and passive attacks. Active attacks are performed by the malicious nodes. It involves modification of data or the interruption of the communication. The passive attacks involve simply monitoring the communication and traffic analysis. It is difficult to detect the passive attack.

ATTACKS IN MOBILE DEVICES

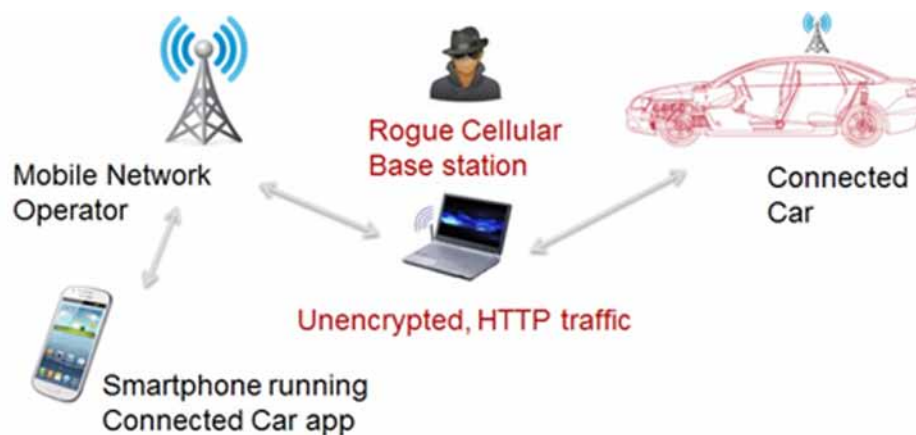
Phishing Attack

The recent years have witnessed that phishing attacks have been increased in the mobile platforms. Mobile phishing is an emerging threat mainly targeting mobile users of financial institutions, online shopping, and social networking companies. This is mainly due to the hardware limitations of mobile devices. It is the type of active attack. Phishing is an identity theft attack online by which the attacker steals the user's credentials resulting in a financial loss as well as a loss to the organization. By the malicious emails, redirecting to the fake websites, impersonation, the attacker can launch the phishing attack. As the mobile device is very compact, and it provides the same functionalities the user moves towards the mobile for all the transaction process. Hence, the attackers targeting mobile devices, for them it is easy to launch the phishing attack on mobile than in desktop systems due to the lack of security systems.

Man-in-the-Middle Attack

This is a kind of active attack in which the attacker sits between the user and the website and steals personal information by breaking the connection between them. Here the attacker impersonates himself as a legitimate user and sends the modified message to the website as it is sent by the legitimate user. This attack is possible even if the message is in an encrypted format. This problem arises when an application fails to use a proper authentication mechanism. Hence, to ensure secure communication, it is important to know how the mobile application performs authentication mechanisms (Mirsky, 2018).

Figure 1. Man-in-the-middle attack



Framing and Clickjacking Attack

Framing is mainly caused by the web browsers like Internet Explorer not checking the destination of the frame. The web framing attacks occur in the browser with the malicious page loading a victim page as an iframe. The common example of framing is clickjacking, where the victim page is loaded which tempts the user to click on that page. This attack enables the wrapper site to execute a clickjacking attack, which is a real threat that has been exploited on high-profile services to steal the credentials of the user or they redirect the users to attacker-controlled sites. The goal of the framing attack is to trick the users into clicking on something different than what they intended. This is mainly to gather confidential information or take control of the affected computer through vulnerabilities. To defend against framing, all the websites use frame busting, which helps to prevent the system from loading a malicious frame in the browser. The new attack called tap-jacking uses the features of the mobile browsers to implement a strong clickjacking attack on mobile phones. The tap jacking illustrates the importance of frame busting on mobile sites. The framing attack can also result in the theft of Wi-Fi passwords or WPA secret key.

Buffer Overflow Attack

The buffer overflow is a common coding mistake by the programmer, which the attacker could exploit to gain access to the system. It is important to understand how the buffer overflow attack occurs and what techniques the attacker uses to successfully exploit these vulnerabilities in the mobile phones.

this attack occurs due to the lack of input validation. The near field communication module of Huawei mobile phones has a buffer overflow vulnerability where the attacker may use an NFC card reader or another device to inject malicious data into a target device. Later Huawei released an update to fix this vulnerability (Bishop, 2012).

Key Concepts

1. This attack occurs when there is more data in the buffer than it can handle; thus, it overwrites the data in the buffer which leads to the loss of information.
2. This attack can even crash the system and create an entry point for a cyber attack.
3. C and C++ are more susceptible to buffer overflow when compared to other programming languages.
4. Secure development or security program practices should include regular testing to detect and fix the buffer overflow attack.

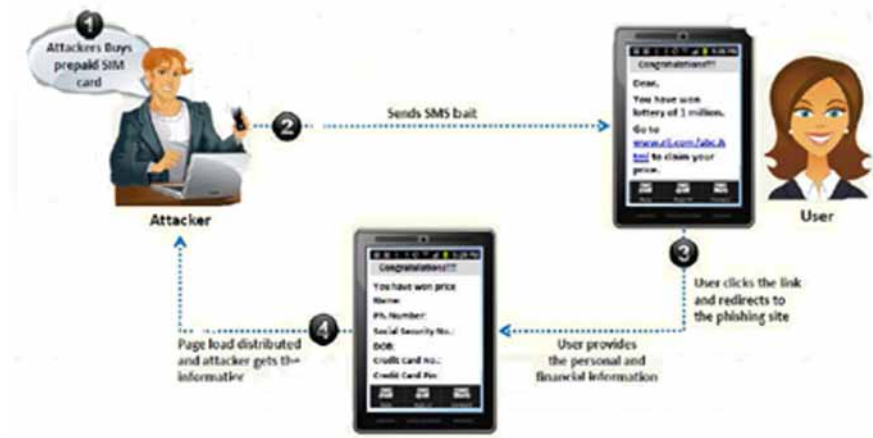
Data Caching

Caching is a technique of storing frequently used data in the memory, such that it could be retrieved easily within a short period. The attack occurs when these data are stolen by the attacker or the attacker can launch the attack by making changes to the cache data which store in the random access memory. Here the attacker replaces the malicious webpage content, code, or some random data he controls to steal the confidential information about the user. Many cache attacks have occurred on Intel X86 CPUs, and the powerful techniques to exploit cache side channels have been developed. However, the modern mobile phones use one or more ARM CPU that have different cache organization and instruction set than in the Intel x86 CPUs. Till now no attack occurs in the non-rooted Android smartphones. But the attacker will try to root the Android smartphone and launch the data caching attack. The routers implement route caching or routing tables to allow faster queries. Then caching in different geographic locations can be seen as content delivery networks (CDN). The CDN is a collection of servers that store local copies of content on servers located strategically across the internet. Each CDN server is a data caching server that provides fast access for the users. Thus, in this way, data caching becomes a major attack that even crashes the server system (Lipp, 2016).

Smishing Attack

This attack is also a security attack similar to the phishing attack but refers to a fraudulent SMS (text message) rather than email. It is a combination of terms of SMS and phishing. Similar to all attacks, the attacker tries to steal personal information. The smishers launch the attack by sending large text messages that are designed to capture the recipient's attention. The messages sent by the attacker provide fake information like "you have won a gift card, visit this webpage to claim it." If the user clicks on this link, then he/she will be directed to the fraudulent website, which will ask you to enter your detail like phone number, email ID. Sometimes it asks for your bank account details. In the smishing attack, the attacker masquerades as a legitimate person to get the user to reveal his/her information. Even the malicious application that is being downloaded from the app store can also cause a smishing attack. The attacker also uses social engineering tactics for convincing the user to do so.

Figure 2. Smishing attack



Improper SSL Validation

The SSL and TLS are widely used protocols for the security purpose. They provide security for the data transmission on the internet between the mobile devices and the remote servers. The SSL/TLS provides end-to-end encryption of the data transmitted, but sometimes it fails to encrypt the sensitive data in which the attack occurs. The mobile devices by nature connect to many different networks and automatically connect to the network. By default it is configured with SSID (service set identifier) for Wi-Fi networks to reduce the traffic. This SSID can be susceptible to man in the middle attack. The basic flow of the SSL handshake is that the client will send a client hello message to the server, and the server will reply will server hello. It is a point in which the certificate validation occurs and also the cryptographic parameter check. In this initialization of SSL itself, there is a possibility of attack that the attacker can steal the secret key shared between client and server. After receiving the server's certificate, the client verifies whether the certificate authority name on the certificate is a trusted CA. If it fails to identify, then the validation fails. Thus, the attacker acts as a legitimate user and uses the services provided by the server (Hinarejos, 2018).

No Encryption/Weak Encryption Algorithm

Different cryptographic algorithms are used in cellular networks. GSM A5/1 algorithm is used for encryption, but it is proved as easily breakable or a weak algorithm. Using eavesdropping itself, the attacker can steal the data. The 3G network provides greater security than the 2G predecessors. It allows the user equipment to authenticate the network and the network will authenticate the device (mutual authentication). 3G network used KASUMI block cipher instead of the older A5/1 stream cipher, but still, several mistakes have been identified in this KASUMI algorithm. Weaknesses in the algorithms used in mobile devices are targeted to the attacks to gather personal information by intercepting the communication.

Privilege Escalation by Mobile Applications

The privilege escalation is an attack of gaining access or privileges to the user's system or a mobile device by exploiting a bug, or configuration in the operating system. The result is that unauthorized

access to the system. Android is a popular open-source mobile platform with prominent features. The weakness in Android's permission mechanism or insecurity leads to privilege escalation attacks. Most of the Android applications are written in Java. They are protected against buffer overflow attacks by bound checking, but can be subjected to attack of a privilege escalation due to the weak security mechanisms. An application with fewer permissions is not restricted to access the components of an application (i.e., it does not ensure the permissions of the device). the privilege escalation can also launch remotely by the attacker, using mail or text message. The iOS platform is also subjected to the same kind of attack as an Android. Mainly in the wireless medium, there will be more possibilities for the attack. Stealing the username and password is also a kind of privilege escalation by which the attacker can compromise a large number of systems. The processing or manipulation of the unknown input leads to this issue.

ATTACKS ON THE NETWORK

Rogue Access Point

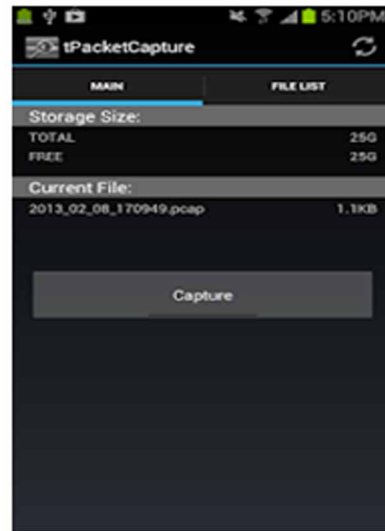
The most common attack in the wireless network is a rogue access point, which is nothing but a device deployed by an attacker to hack the information of the users. It can easily be deployed but is difficult to detect. Almost every mobile using the internet is using Wi-Fi, which is nothing but an access point. Thus, the attacker will place a rogue access point (fake access point) with high frequency and strong signal to connect with the user, and as soon as the user connects with a fake wireless access point (WAP), he gets trapped. It may even cause a denial of service attack, malicious data insertion, and data destruction. With this rogue access point, the attacker can do an active and passive attack that we have seen before. According to the report of AirTight out of total access point, 20% are unauthorized in the network, and the users will easily connect to it due to the lack of knowledge about the security threats (Sadip, 2013).

Packet Sniffing

Sniffing is a process of monitoring the activities of a system in the network. When our system is connected to the network, many transmissions will occur. The sniffer, a kind application, can read, monitor, and capture data exchanges and read the data packets. By capturing these packets, the attacker can gather a lot of information about the user by which the attacker can launch the attack. The mobile phones are the major target for the attacker due to the increase in the number of mobile users, and all the users are accessing all the applications in the device itself because of their scalability and convenience. Even the payment process through banking application is also done via mobile devices. Hence, the attacker targets mobile devices rather than the laptop or system, by which the attacker can gather more sensitive information and even the bank account details. All the packet sniffer tools will work only on the rooted mobile devices; therefore, the attacker will first try to root the device before starting to capture the packets. It is also possible to sniff the cellular communication traffic by the attacker, thereby makes the service not available to the users and steal the identity (Mao, 2017).

The tools for packet sniffing in the mobile devices of the android platform are packet capture, FaceNiff, Android Pcap, etc.

Figure 3. tpacket capture tool



Session Hijacking

Session hijacking is an attack of stealing one's valid session or a token to gain unauthorized access to the services in the system. Session hijacking affects the integrity of the data on a network. A network with dynamic nodes or no central administration leads to network attack in which one of them is session hijacking. The combination of a man in the middle attack and denial of service attack is used to launch session hijacking. In session hijacking, the attacker exploits a web session control mechanism, which is normally managed by a session token. The attacker compromises the session token by stealing or predicting a valid session token to gain access. They can steal the session by session sniffing, client-side attacks, and man-in-the-middle attacks. The attacker will predict the session by identifying the pattern of the session token which they gather by sniffing.

DNS Poisoning

The DNS works in such a way that, if the browser needs to visit any website like www.security.com, it needs to find the IP address of the webserver. First, it asks the system's host's file, which may have the IP address of the DNS if it is visited already. Then it asks the external ISP (Internet Service Provider). Normally if we have already visited this site the system has the IP address in its cache itself. Sometimes it will be expired. Here the attacker will steal the cache of the browser and he can change the IP address of the particular DNS, thereby redirecting to the attacker's fake website. Sometimes this attack will lead to the denial of service. The attacker will modify the information in the cache of a server that is part of the domain name system, which leads to a cache poisoning attack. The DNS information has time to live associated with it. To overcome this DNS poisoning, the TTL for DNS information can be specified by the DNS owner to retrieve the DNS information in a short period (Kim, 2011).

SSLStrip

The SSL strip is a type of man in the middle attack in which the website secured with HTTPS is downgraded to HTTP. In SSL strip all the traffic coming from the user's machine is directed towards a proxy website which is created by the attacker. This attack can be done in several ways. Very common methods among them are manually setting up the proxy of the browser to route all the traffic. Then ARP Poisoning, for example, assumes that there is a private connection between the user and the server of a bank, where all the traffic is transferred through a secure tunnel of SSL which has a URL of `https://www.onlinesbi.com`. At this point, the attacker modifies the response from the server from `https` to `HTTP` and sends it to the legitimate user with the URL of `http://www.onlinesbi.com`. Then the attacker can easily gain access to the account, and all the information is transferred from user to server as plaintext as there is no secure connection between them. Hence, the attacker can steal all the sensitive information.

ATTACKS ON THE WEBSERVER

Cross-Site Request Forgery

It is an attack that forces the end-user to perform a malicious action by a click on the website. Thus, the attacker steals the sensitive information of the user. The attacker manipulates the scripts to the web applications to execute based on his/her needs to gather the victim's information. By this attack, the attacker can hijack user sessions, deface a website, and redirect the user to malicious websites. It happens mostly in the web application when the clients have access to the cookies or other authentication mechanisms. The mobile application has a web viewer that will have cookies for its system. The cookies stored in this system can be shared with the applications which lead to a threat of an attack. In mobile platforms especially in Android and iOS, the application will not share their cookies with the other applications. Hence, the possibility of an attack is lesser. It is not likely within the mobile application. The attacker exploits the authentication of the web applications.

Weak Input Validation

Due to the increased users of the internet, many web applications are becoming targets for attack, in which the most common attack is weak input validation. Any web application should satisfy its security features at least to some extent. The application with the script of weak input validation will allow the attacker to gather more information without the knowledge of the admin. The attacker will give any input in the webpage irrespective of the correct value, for example in the phone number field, which should be designed in such a way that it should accept only the integer value, but here the attacker will enter a string. Hence, it produces an error which reveals some information if the validation is not done properly. In the development stage itself, the developer paid only a little attention to the security part. If someone breaks the security of the web application, then only they will think about the security. This kind of behavior of developers is also a threat to this attack, especially in web page design.

Server Misconfiguration

Server configuration is an essential part of web applications. The server misconfiguration attacks exploit the weakness in the web application. Thereby the attacker can bypass the authentication mechanism itself. Enabling the unnecessary services, giving access to the unwanted things, sometimes provides access to the script of the web application and all the exploits of the server misconfiguration attack. All the web applications will have some vulnerabilities due to the server misconfiguration by which the attacker will gain access privileges. Now the attacker has total control the web application he can do an attack on the particular web application.

ATTACKS IN THE DATABASE

SQL Injection

The SQL injection is a malicious code injection attack in the database to steal the information or to crash the database. The attacker uses the input field to steal the data or to gain unauthorized access. It occurs due to the non-validation of the input field. In Android mobile devices, SQLite is used, which is fully functional databases. Here also the attack occurs similar to SQL Server or MySQL. The root cause of the SQL injection is the use of dynamic queries. It is a flaw in the web applications, not in the database or the webserver. Based on the type of application and the way of processing data, the SQL injection can be used to implement attacks like authentication bypass, information disclosure, remote code execution.

Privilege Escalation Attack

Privilege escalation is an attack by which the attacker gains access to mobile devices. It occurs by exploiting the bug or design flaws and weaknesses in security implementation. There are two types of privilege escalation. They are horizontal and vertical. In horizontal privilege escalation, the attacker targets the normal end-user, where the attacker doesn't gain any advanced privileges and simply steals one's identity to gain access. Then in vertical privilege escalation, the attacker gains advanced privileges access to the user account.

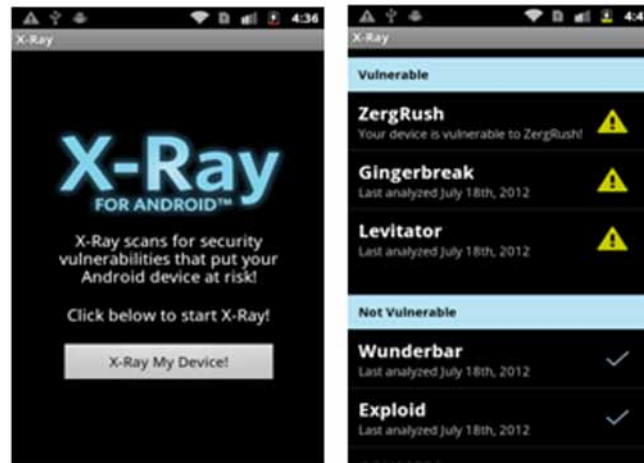
DETECTION MECHANISM

All the applications in the mobile will have some vulnerabilities. They should be detected as early as possible to overcome the security issue. Many vulnerability scanners are used in the mobile device that scans for the vulnerability and threats and takes preventive measures. X-ray is a vulnerability scanner tool for the Android platform, and also by analysis methods, we can detect the threats they are static analysis, dynamic analysis, and permission-based analysis (Goel, 2018).

X-Ray Tool

The x-ray tool scans your mobile device to detect vulnerabilities that remain unpatched. After scanning it lists all the vulnerabilities identified in the system and checks for preventive measures. This tool will automatically update to scan for new vulnerabilities.

Figure 4. X-ray scanning tool



Static Analysis

In the static analysis method, the properties of the applications and their source code are inspected or analyzed to detect the threats. However, it is difficult in case of, obfuscation and encrypted techniques embedded in the software. Further, the static analysis is classified into two categories they are behavior-based and signature-based on antivirus.

Dynamic Analysis

The dynamic analysis involves the execution of the application in an isolated environment to track its behavior. In this method, it discloses the normal behavior of malware as executed code. Hence, it is immune to the obfuscation. For suspicious application detection in the Android platform, both static and dynamic analyses should be done. In dynamic analysis, there is also another method called run-time-based behavior dynamic analysis system for Android applications. This system consists of an event detector log monitor and parser. With this collected information, we can detect malicious applications. The tool for dynamic analysis TaintDroid used for system-wide android application by tracking the flow of sensitive information via third-party applications.

Permission-Based Analysis

The permissions requested by the mobile application plays a vital role in providing the access rights. By default, the application won't have any permission to access the data on the mobile. As a user when

we are installing the particular application, allowing that to access the resources. With the help of the listed permission of the applications, we can able to detect the malicious behavior of the application. These permissions can limit application behavior by controlling our privacy. With the static analysis itself, the required permissions can be obtained based on the functionality of the applications such that the insecure applications have been detected by this method of analysis.

PREVENTION MECHANISMS

It is the first line of defense against the attack or a vulnerability in any system. Some of the security guidelines to provide or ensure security for the mobile devices are as follows (Dua, 2014):

- We should do not allow too many applications to install automatically and do not load a large number of applications in the limited memory space and avoid auto-upload of photos to social networks.
- It should perform security and vulnerability assessment of the application architecture.
- Always maintain configuration control and management of the application on the mobile of different platforms.
- The applications should be installed from the trusted application stores.
- In the case of disposing of the device, the data in the device should be securely wiped or deleted.
- Ensure that the Bluetooth and wifi are all turned off by default. And the mobile system should be in such a way that the user should have permission to turn on or off based on their needs.
- Never connect two different networks simultaneously (i.e., wifi and Bluetooth)

Then the user should provide some security mechanisms to defend against the attack they are (Bitton, 2018):

1). Using Password or Passcode:

By configuring a strong password of maximum possible length, which should contain letters, numbers, and special characters for the mobile device, we can provide security for the device. Then set an idle timeout to lock the phone automatically when not in user and enable lockout/wipe feature after a certain number of attempts.

2). Enable Remote Management:

In the enterprise environment, user Mobile Device Management (MDM) software secures, monitors, manages, and supports mobile devices deployed across the organization.

3). Use Remote Wipe Services:

Locating the mobile device or to identify the device at any location by using remote wipe services. It will be helpful when the device is lost or stolen.

4). Update OS and Apps:

Update the operating system whenever the update is available, by which we can reduce the risk of attack. All the applications in the mobile device installed should be updated.

5). Do Not Allow Rooting or Jailbreaking:

Rooting the device will bypass all the security mechanisms, hence rooting, should not be done in the mobile device. Should include this as a mobile security policy.

6). Encrypt Storage:

If there is any mechanism supported to encrypt the storage, configure your mobile device to encrypt its storage with the hardware encryption.

7). Perform Periodic Backup and Synchronization:

The backup should be taken from the mobile device to prevent the loss of data and perform periodic synchronization.

The mobile security policies should be implemented to prevent it from the attack. Then harden the browser permission rules and configure application certification rules, by which mobile devices can be prevented by the attacks. Then some general guidelines for the mobile platform security include:

- Configuring wireless to ask to join networks
- Perform regular software maintenance
- Control the location of backups
- Control device and applications
- Prohibit USB keys
- Encrypt backups
- Prevent local caching of email
- Sandbox application and data

Mobile tracking tools for iOS platform include finding my iPhone, iHound, GadgetTrak iOS security, iLocalis. Then for the Android platform the tools include find my phone, prey, my antitheft, where's my droid.

The mobile security guidelines for the administrator:

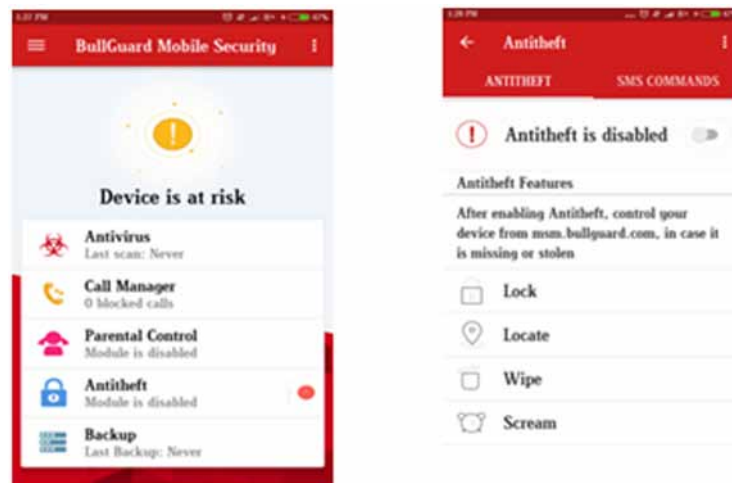
- Publish an enterprise policy that specifies the acceptable usage of consumer-grade devices and brings your devices in the enterprise.
- Publish an enterprise policy for the cloud.
- Enable security measures such as antivirus to protect the data in the database.
- Specify a session timeout through the access gateway.
- Specification of the password is also considered by the admin whether the domain password can be cached on the device, or the user must enter it when a request for access.

MOBILE PROTECTION TOOLS

Bullguard Mobile Security Tool

Acts as an antivirus tool and scans for risks in the mobile device and provides security from the attacks.

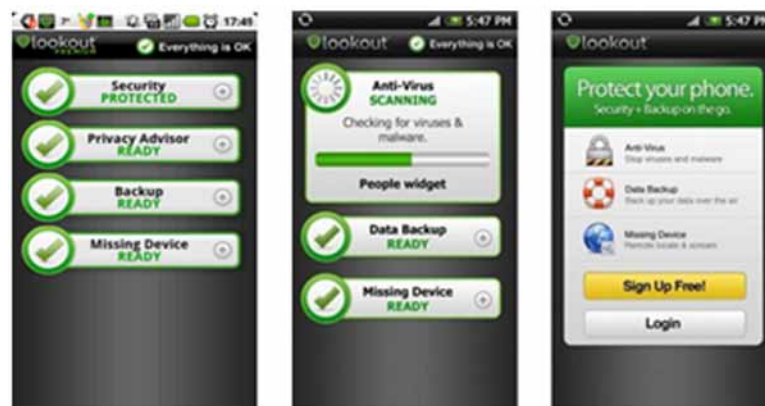
Figure 5. Bull guard protection tool



Lookout

- It protects the phone from mobile threats.
- Helps in avoiding risky behavior, like connecting to an unsecured wifi network, downloading a malicious app.
- It provides a secure backup of your mobile data.
- It helps you to find your device if it lost or stolen.
- Provides remote management.

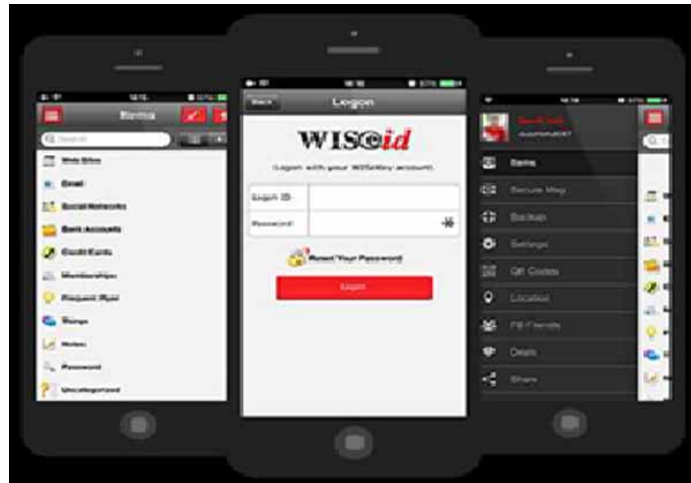
Figure 6. lookout protection tool



WISelD

- It provides secure and easy to use encrypted storage for personal data.
- It allows the user to store websites, usernames, and passwords and quickly log on to the particular website through your mobile phone.

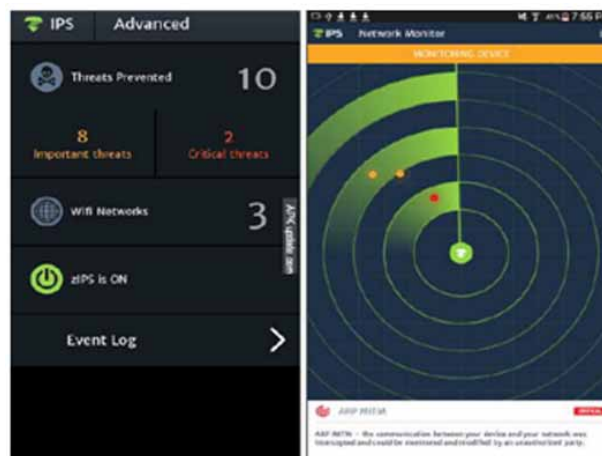
Figure 7. wiseid mobile protection tool



zIPS

- Employs machine learning to detect abnormal behavior and isolate your device before any exploit can take place.
- It is equipped with a behavioral analysis engine to automatically detect and block malicious threats by monitoring how they change the characters of the mobile device.
- It scans all the mobile applications and browsers to enhance the security of the user devices and keeps your whole organization safe from MITM, IPv4, and even IPv6 attacks.

Figure 8. zips tool



Firewall and Intrusion Detection System

With the help of firewall and IDS software, we can prevent the mobile device from the attacks. The mobile device is a wireless device hence there exists a higher possibility of an attack when compared with the desktop system. It is better to use a firewall and IDS system to ensure security for mobile devices.

CONCLUSION

The smartphone is becoming more popular in terms of communication, sensor, and for doing any activities online, we can say it replaces all the desktop and laptop systems due to its scalability and portability. Even though it is very small in size, it provides a lot of services such as messaging, browsing, emailing, playing games, and payments. As the services increase in the mobile device, the possibility of the attack also increases. Many new threats, vulnerabilities, and attacks are increased due to its multi-functionality. Here we have discussed the threats and vulnerabilities of the mobile applications, and also the attacks that occur from these threats. Finally, we have discussed the detection and prevention mechanisms and the mobile security tools that are used to protect the mobile device.

FUTURE RESEARCH DIRECTIONS

This chapter provides a detailed description of all the threats and vulnerabilities of the mobile applications that we are using in our day-to-day life. Hence, with this research, we can proceed further to improve the present vulnerable mobile application with the prevention and detection mechanisms and also by providing guidelines for the developers. This may help people to reduce the threat to some extent. Several research works can be done more than this, like the countermeasure for the specific threat to improve the present situation by implementing some new methods. More threats and vulnerabilities other than discussed and can be analyzed to find the countermeasure in the future, such that it may not occur again.

REFERENCES

- Bishop, M., Engle, S., Howard, D., & Whalen, S. (2012). A Taxonomy of Buffer Overflow Characteristics. *IEEE Transactions on Dependable and Secure Computing*, 9(3), 305–317. doi:10.1109/TDSC.2012.10
- Bitton, R., Finkelshtein, A., Sidi, L., Puzis, R., Rokach, L., & Shabtai, A. (2018). Taxonomy of mobile users' security awareness. *Computers & Security*, 73, 266–293. doi:10.1016/j.cose.2017.10.015
- Delac, Silic, & Krolo. (2017). *Emerging Security Threats for Mobile Platforms*. Academic Press.
- Dua & Bansal. (2014). Review on Mobile threats and Detection techniques. *International Journal of Distributed and Parallel Systems*.
- Goel & Jain. (2018). *Mobile Phishing Attacks and Defence Mechanisms: The State of Art and Open Research Challenges*. Science Direct – Computers and Security.
- He, D., Chan, S., & Guizani, M. (2015). Mobile application security: Malware threats and defenses. *IEEE Wireless Communications*, 22(1), 138–144. doi:10.1109/MWC.2015.7054729

- Hinarejos, M. F., Almenarez, F., Arias Cabarcos, P., Ferrer-Gomila, J. L., & Lopez, A. M. (2018). RiskLaine: A Probabilistic Approach for Assessing Risk in Certificate-Based Security. *IEEE Transactions on Information Forensics and Security*, 13(8), 1975–1988. doi:10.1109/TIFS.2018.2807788
- Khan, J., Abbas, H., & Al-Muhtadi, J. (2015). Survey on Mobile User's Data Privacy Threats and Defense Mechanisms. *Procedia Computer Science*, 56, 376–383. doi:10.1016/j.procs.2015.07.223
- Kim, H., & Huh, J. H. (2011). Detecting DNS-poisoning-based phishing attacks from their network performance characteristics. *Electronics Letters*, 47(11), 656. doi:10.1049/el.2011.0399
- Lipp, Gruss, & Spreitzer. (2016). *AR Mageddon: Cache Attacks on Mobile Devices*. Academic Press.
- Mao, J., Tian, W., Li, P., Wei, T., & Liang, Z. (2017). Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity. *IEEE Access: Practical Innovations, Open Solutions*, 5, 17020–17030. doi:10.1109/ACCESS.2017.2743528
- Mirsky, Y., Kalbo, N., Elovici, Y., & Shabtai, A. (2018). Vesper: Using Echo-Analysis to Detect Man-in-the-Middle Attacks in LANs. *IEEE Transactions on Information Forensics and Security*.
- Razzaq, A., Hur, A., Ahmad, H. F., & Masood, M. (2013). Cybersecurity: Threats, reasons, challenges, methodologies, and state of the art solutions for industrial applications. *2013 IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS)*. 10.1109/ISADS.2013.6513420
- Seo. (2011). *Kangbin Yim*. Mobile Malware Threats and Defenses for Homeland Security.
- Sheik Abdullah, A., & Priyadharshini, P. (2020). Big Data and Analytics. In A. Haldorai & A. Ramu (Eds.), *Big Data Analytics for Sustainable Computing* (pp. 47–65). IGI Global. doi:10.4018/978-1-5225-9750-6.ch003
- Thite, & Vanjale, & Mane. (2013). Elimination of Rogue access point in Wireless Network. *International Journal of Scientific and Engineering Research*, 4(12).
- Xenakis & Ntantogian. (2015). *Attacking the Baseband Modem of Mobile Phones to Breach the Users' Privacy and Network Security*. Academic Press.

KEY TERMS AND DEFINITIONS

Attacks: An offensive action that targets the system networks, infrastructures, or personal computer devices to steal sensitive information.

Authentication: Process of checking the identity of a person or a device.

Encryption: Process of encoding the information in a different format that can be understood only to authorized persons.

Exploits: Refers to a piece of software that takes advantage of the vulnerability, thereby damaging the whole process of the system. It can occur in both the software and the hardware.

Malware: It is malicious software that can be in any file or a program.

Obfuscation: Deliberate act of creating a machine or a source code that is not understandable for humans.

Ransomware: A type of malware that can deny the accessibility of the system until the ransom is paid.

Security: Refers to the activity of protection for the computer system in terms of usability and integrity. It includes both the hardware and the software.

Spyware: An unwanted software that has the ability of stealing the sensitive data, without your knowledge.

Threat: The uninterpreted event that causes a risk that leads to loss.

Trojan Horses: Refers to a type of malware that seems to be a legitimate application.

Viruses: Malicious code that replicates itself to another program, file, or document, thereby changing the process in the system.

Vulnerability: Refers to a flaw in the computer system, which creates a way open to the attack.

Worms: Similar to viruses, but it replicates itself to spread to the other systems connected to the same network.

Web Authorization Protocols

4

Demetrios Georgios Syropoulos-Harissis

Greek Molecular Computing Group, Greece

Apostolos Syropoulos

 <https://orcid.org/0000-0002-9625-1482>

Greek Molecular Computing Group, Greece

BACKGROUND

Web applications were designed so to be a simple and friendly way for humans to interact with a machine. Initially, they were custom made and could solve a few problems. However, as the technology evolved, it was possible to create Web applications that could solve an even bigger range of problems. Nowadays, Web applications can virtually do almost anything. Apps are stand-alone applications that implement the full functionality of a Web page/application (Ater, 2017). These apps are typically installed on a mobile device (e.g., a smartphone or a tablet). There are simple but useful applications (e.g., apps that provide information about the weather) and more complicated ones (e.g., the Facebook and the LinkedIn apps). Most of these apps have something in common: They require some sort of account creation in order to personalize the data that they will offer to the user. Of course, this is a major development for apps. In the past, one could select a city and see the weather forecast in detail. One question that comes into our mind is: Why has the weather forecast has been so personalized and why Facebook needs to know my location at any given moment?

Most people are very sensitive about their personal data. Not all people consider the same kind of data, personal. Nevertheless, there are data that are definitely personal. For example, one's location or personal messages are definitely personal data. Apps require some of these personal data in order to provide a fully personalized experience. But this is something that makes people very suspicious. There are two problems here:

- i. How is this information going to be used?
- ii. How can be sure this information will not end up in wrong hands?

Fortunately or unfortunately, in the end, most users give access to their personal data. People do that because they trust their information to apps and whoever is behind them (e.g., a company, a developer, etc.). Typically, people who have “adopted” *stereotypes* (Neil Macrae, Stangor, & Hewstone, 1996), assume that developers have taken every precaution so that an app uses the data in a safe and sound way. Naturally, this is not always the case... Although the data management problem has been relatively easily solved in the past, still there are other more complicated problems that have come up after it. With so many personalized apps, anyone has typically an account for each app that she is using. If someone wants to use n apps, then that person needs n different accounts! [This is a major problem that was examined quite early, for example see (Adams & Sasse, 1999).] Having many accounts implies the use of many *authentication* systems. An authentication system makes sure that when a user attempts to login to a system or an app, she will be granted access only when she enters a valid user name and the

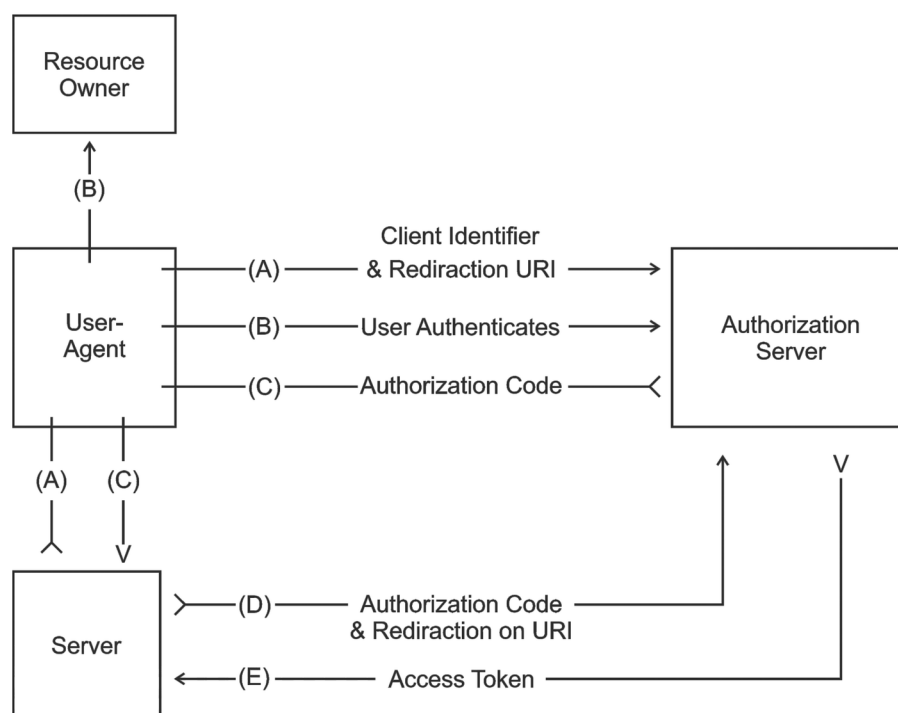
DOI: 10.4018/978-1-7998-3479-3.ch035

corresponding (secret) password. After user authentication, a system proceeds with *user authorization*, which is a process that verifies that the right person has access to the right data. A simple example for the discrimination of authentication and authorization is the following. Consider a student who enters the library. She shows her membership card to the security officer and enters. This procedure can be thought of as an authentication process. Next, the student is searching for a book. She finds the book and takes it to the reading room that is guarded by someone. Our student is asked for her membership card, but her *credentials* are not adequate as the reading room is for teachers only. The second check can be thought of as an authorization procedure. So the problem with many accounts and different authorization systems for a single app is solved with the *authorization protocols*. These protocols give us the ability to use one, two, or more accounts to connect to any number of applications in a secure and safe way. In this article we give an overview of authorization protocols and safe data transfer between apps.

INTRODUCTION

When we hear the term “protocol” what comes first in mind is a number of rules someone has to follow in a certain occasion. These rules have been given by someone who created this protocol and it is considered that these rules, if followed, are the best way to deal with this occasion. For example, think of banquet protocol and how one has to arrive at such an event, how she is supposed to dine, etc. An authorization protocol is definitely a protocol but differs slightly with what we described so far. In particular, a protocol is supposed to be unchanged and strict. However, the authorization protocols are being upgraded every time some security error occurs. We have introduced two new very important terms now: *upgrade* and *security*. The first term, upgrade, shows the flexibility of these protocols and it is being implemented

Figure 1. Authentication process. PR is a reliable partner and OP is an OpenID provider



with new protocols based on the old ones or layers that work along with the protocol. The second term, security, is what it is all about. When someone grants access to an app to his personal data, then one is mainly concerned about the integrity and the security of her data. For example, if a very personal photograph (e.g., one that shows an intimate moment) suddenly becomes public, then this is a security error. Thus we must ensure that these protocols are first of all secure.

Our effort to solve the problem with multiple accounts when using many different apps made it clear that we need authorization protocols. Nowadays, things have changed and we typically have one or maybe two accounts in different Web pages/apps and manage to connect to the rest of our accounts using these two apps. This means that we allow these apps to exchange information and data with the rest of the apps. This is realized in a secure and safe way because the way this exchange happens is controlled by the authorization protocol. This protocol guarantees that the correct information from the correct user will be shared in a safe and secure way with the correct app. In a way, this description is an oversimplification, however, it describes the main idea of authorization protocols and also explains how it works with many and different accounts.

OAuth: An Authorization Protocol

When we create a process in which sensitive data are being exchanged and/or shared, we have to make sure that this is a safe process and take every necessary precaution so not to get compromised. This is the ultimate goal for any authorization protocol, in general, and for OAuth, in particular. From the first moment the OAuth (OAuth Community Site) protocol made publicly available, it has been used widely in software engineering and by computer software companies. Currently, there are two versions. Let us now explain how, at least in theory, OAuth operates. But first, let us introduce some simple terms related to Web programming.

- **Server:** It is a computer program that provides certain Web services to other programs or devices (e.g., think of a Web-mail server).
- **Client:** It is a computer program that is able to use some services provided by a server (e.g., a Web-mail client).
- **End-User:** Typically this is a (real) person that uses a device, a Web page, or a stand-alone computer program.
- **Access Token:** It is a sequence of characters considered as an entity that is used get access to a database or information provided by a server. An access token is not forever valid and expires after some time. When an access token expires, it can be replaced with the *refresh token*.

We are sure that most readers have the opportunity to try to login to a Web page or an app using the credentials they are using for a different Web page or app. Since most people have either Facebook and/or Twitter accounts, when they want to use a new app (e.g., a fitness app), they will be asked to sign up using their Facebook or Twitter account. Of course, users can always choose to create new accounts with new credentials. Suppose that someone chooses to sign in using his Facebook account. Then, the new app will exchange some data with Facebook's site and presto the user is granted access to the new app. What really happened beneath is that the new app has acquired information about the user and is using them to create a new user profile. In this really short time, the client (e.g., the fitness app) and the server (e.g., the Facebook site) are following the OAuth (or OAuth 2.0) protocol in order for the client to obtain from the server the required information. Both the Facebook site and Google mail site are ex-

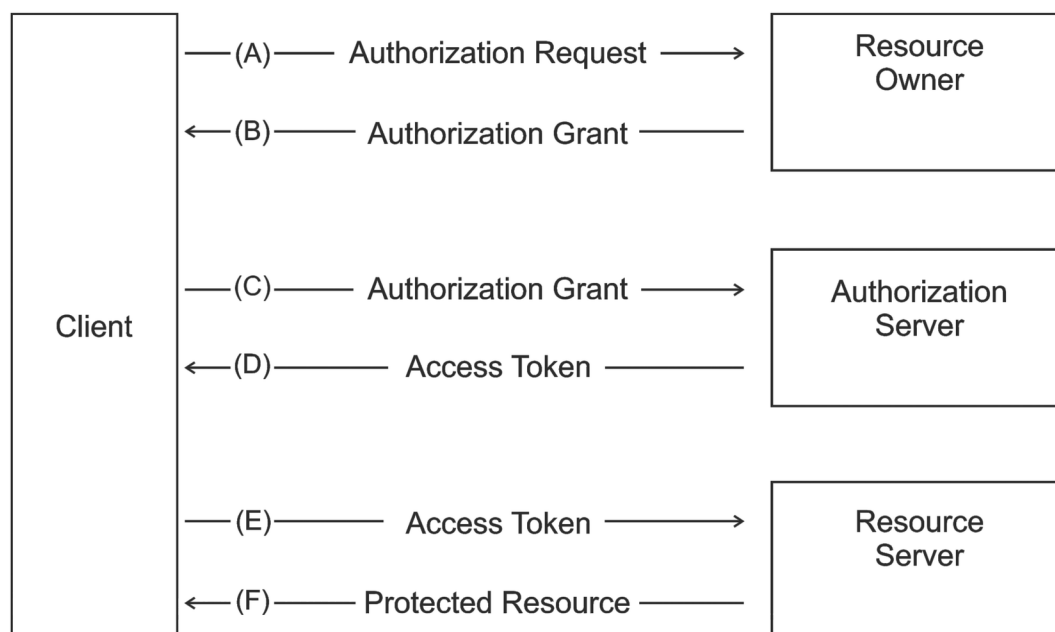
amples of sites that have used OAuth and now use OAuth 2.0 (Hardt, 2012) in a compatible way. OAuth compatibility is a very important aspect of the authorization process. It ensures that both the client and the server have the necessary tools installed to proceed with the authorization process.

OAuth as an authorization protocol had some great advantages and some security errors. Naturally, the advantages remained in the next generation of the protocol. In addition, most security errors have been fixed in the new version. One of the strong characteristics of OAuth is that each information exchange session is considered unique and since certain security measures are in effect, it is also secure. At this point it is rather important to say that OAuth was not as widely used as the OAuth 2.0 because of the limited demand at that time (it appeared in 2006). Also, the next version, OAuth 2.0, was released a short time after OAuth and it was much better than OAuth. However, since OAuth 2.0 was so widely adopted, many different types of authorization code flow were designed. These different types have been designed so to solve the speed problem (sometimes authorization was painfully slow). They also tried to make instances that work for specific data exchanges that are not so sensitive. Nevertheless, OAuth 2.0 main *grant* (in OAuth parlance types are called grants) is one and it is used in more than 70% of cases. Why? Simply because it is the most secure. This grant is the *authorization code grant* (Authorization Code Grant).

Authorization Code Grant

The main component and most important aspect of this grant is the authorization code. In general, this is a sequence of characters that encode certain kinds of information. It is also unique for every session. These sequences of characters include information about the server and the client and kind of information can be exchanged. This binds them both to an exchange that is safe. The authorization flow follows these steps (see Figure 2):

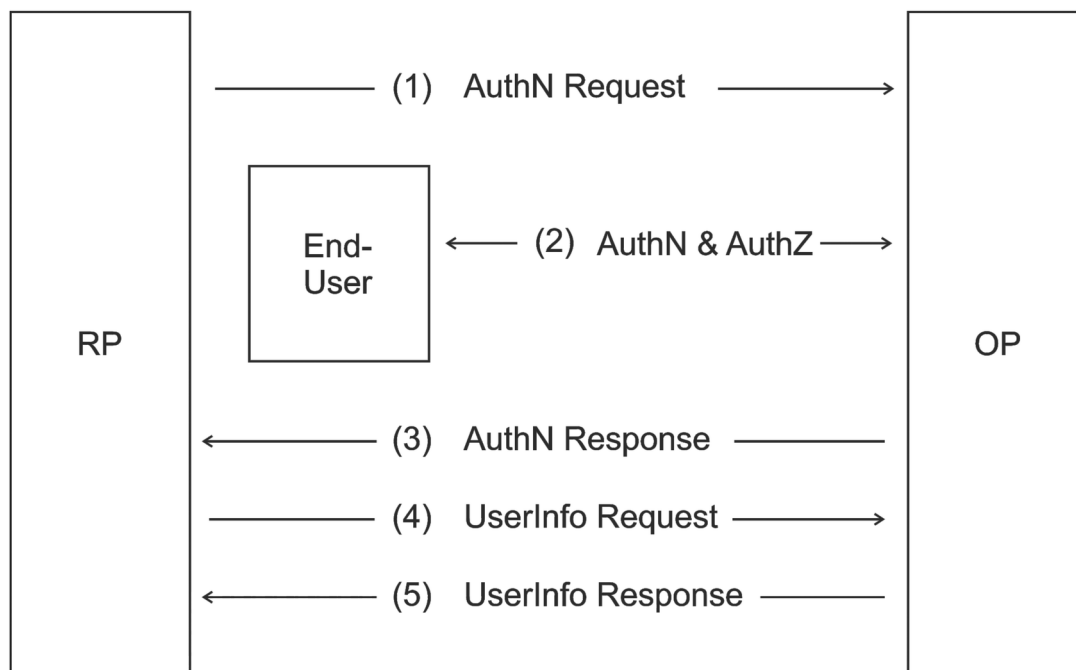
Figure 2. Abstract Protocol Flow



1. The client gives to the end-user the necessary information in order to proceed with the flow and redirects it to the server in order to authenticate.
2. The End-user authenticates herself with a login page provided by the server.
3. If the End-user authenticates successfully, she is asked if she allows the client to proceed with the information exchange.
4. If she agrees with the information exchange, the server creates an authorization code based on the information that have been given by the client and the information that will be exchanged. The server redirects the End-User back to the client with the authorization code.
5. The client uses the code to authenticate with the server and asks for an access token.
6. The server creates and gives the access token to the client if the authentication process is complete.
7. With this access token the client can now access the information.

From the description of the operation flow of the OAuth 2.0, it should be clear that the authorization code is a secure flow (see Figure 3). It should be also clear that each session is unique as is the exchange of specific information (i.e., one can get the same information from another session).

Figure 3. Authorization Code Flow. Note: The lines illustrating steps (A), (B), and (C) are broken into two parts as they pass through the user-agent



OAuth 2.0 Evolution: Open ID connect

As the use of OAuth 2.0 became widely accepted, many security issues came up. Most of those are not obvious. They can be exploited but this is not a particularly easy task. Common security problems in OAuth 2.0 are happening when the authorization code of session is released or when someone can somehow find a way to obtain these codes. These problems can be solved when the tools that we have used in an

implementation of OAuth 2.0 are the best possible. Unfortunately, in many cases these tools are not used. The reason for this is that it takes time to handle all possible security problems. And by time we mean the time it takes to implement all these additional checks and the fact that these additional checks will make an application quite slow. Thus people always make compromises in order to get an optimal result.

To fix the problem that occurs because of the security issues a new addition to the OAuth 2.0 was created, the Open ID connect (Open ID connect, 2019). This is a layer that works on top of the protocol and fixes the security issues that were created.

Open ID connect or just Open ID is a layer that works on top of OAuth and has as its main component the ID token. This token works in a way that is similar to the way the authorization code works only it is much better. It is a JWT type, where JWT stands for *JSON Web Token*. This is a JSON type token. JSON, the Javascript Object Notation, is an open standard file type that is used for data transfer between a server and other components. The JWT has a certain structure that is composed by three components:

1. The header
2. The payload
3. The signature

Each of these components is created in every session and they are unique, safe and secret. The header contains two pieces of information that are necessary to every JWT and determine the use of the JWT. The payload contains the information that we need to use for authentication and authorization. The signature is the header and the payload is encrypted. Afterwards, the signature is used to encrypt the whole JWT again. The extensive encryption that takes place in the creation of every JWT makes the process reliable and secure. The encryption methods that are used vary and there is no requirement as to which method to use.

The great advantage of using a JWT is that someone can include many pieces of information inside. And this happens because it was opted to use JSON to encode JWTs. The ID token contains information about the client and its authentication. It also contains information about the authentication of the end-user. The information always arrive safe and sound to the server. From there the server can perform the authorization and proceed with the process.

With the use of the Open ID connect all the data exchanges became safe and secure. However, there is a little price to pay because of the process time: encryptions and decryptions are time consuming processes. But the ultimate goal of a secure and safe session is guaranteed so this is the best solution.

FUTURE RESEARCH DIRECTIONS

There are always new challenges one has to face and/or solve when dealing with security and/or authorization. For example, a blogger has noted that “[OAuth 2.0 is] overly prescriptive and, given that there isn’t a single conformant implementation by a major service provider, perhaps too complex for even the big boys” (OAuth 3.0: The Sane and Simple Way To Do It). And this is the reason why people are working on the next version of this protocol. Another similar problem is the use of web banking. People are forced to have an unique web account for the bank accounts they keep in a specific bank. However, if someone has two or even three accounts in two or three banks, then she obviously needs three different web accounts. Clearly, it would be nice to allow users to login with a single account, but of course there many problems that have to be solved before this may happen.

CONCLUSION

Data exchange while using applications is very common nowadays. Securing these processes by making reliably protocols is a top priority. The user will never be able to fully understand how these protocols actually work in the background. Some developers make sure that these protocols are not only secure but they are using only the allowed information and distributing them correctly and safely. There will always be security errors in these processes. What developers are trying to achieve is to minimize the risk of a big data leakage in order to provide reliability and trust to the user.

REFERENCES

- Adams, A., & Sasse, M. A. (1999). Users Are Not The Enemy. *Communications of the ACM*, 42(12), 41–46. doi:10.1145/322796.322806
- Ater, T. (2017). *Building Progressive Web Apps*. O'Reilly Media, Inc.
- Authorization Code Grant. (n.d.). Retrieved 05 30, 2019, from <https://www.oauth.com/oauth2-servers/server-side-apps/authorization-code/>
- Hardt, D. (2012, October). *Internet Engineering Task Force*. Retrieved August 25, 2019, from The OAuth 2.0 Authorization Framework: <https://tools.ietf.org/html/rfc6749>
- Neil Macrae, C. N., Stangor, C., & Hewstone, M. (Eds.). (1996). *Stereotypes and Stereotyping*. Guilford Press.
- OAuth 3.0: The Sane and Simple Way To Do It. (n.d.). Retrieved 05 30, 2019, from Tav's Blog: <http://tav.espians.com/oauth-3.0-the-sane-and-simple-way-to-do-it.html>

KEY TERMS AND DEFINITIONS

Exploit: A piece of software that takes advantage of vulnerabilities in applications, networks, or hardware.

JSON (JavaScript Object Notation): Is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate.

OAuth: A simple way to publish and interact with protected data.

Open ID Connect: Is a simple identity layer on top of the OAuth 2.0 protocol.

Secure and Reliable Knowledge-Based Intrusion Detection Using Mobile Base Stations in Smart Environments

Ambika N.

 <https://orcid.org/0000-0003-4452-5514>

Department of Computer Applications, SSMRV College, Bangalore, India

INTRODUCTION

Wireless sensor network (Yick, Jennifer, Mukherjee, & Ghosal, 2008) (Akyildiz, F., Su, Sankarasubramaniam, & Cayirci, 2002) is an assembly of low cost nodes which can capable of self-configuration. These nodes communicate with each other forming a convenient topology. Smart home (Erol-Kantarci, Melike, & Mouftah, 2011), military surveillance (Lee, Hyuk, Lee, Song, & Lee, 2009), health monitoring (Kim, et al., 2007), habitat monitoring (Mainwaring, et al., 2002) are some of the applications where these nodes play a primary role. These unsupervised nodes are liable to get compromised and launch different kinds of attacks. Hence the malicious nodes have to be detected (Butun, Ismail, Morgera, & Sankar, 2014) at an early stage to minimize the attacks in the network.

The mobile base station (Sun, Osborne, Xiao, & Guizani, 2007) (Djamel, Khelladi, & Badache, 2005) provides flexibility to the system to move to any corner of the network. They are capable of authenticating the other nodes but they require more security than the static ones. The proposed work is using the mobile base station that adds reliability to the data transmitted. The work is compared with (mehmood, et al., 2018). The work considers two different kinds of nodes in the network. Assistance nodes are the watchdogs in the work. They aid in selection of cluster head and provide information on any malicious activity in the respective cluster. The mobile agent is used in the work to cross-verify the malicious activity of the node deployed in the cluster.

The work is divided into six sections. The section 2 details the work done by various authors. The section 3 explains the proposed work in detail. The role of different deployed nodes is described in this section. Section 4 provides details of analysis of the work. The details of simulation of the work are provided in section 5. Section 6 details the future directions. The work is concluded in section 7.

BACKGROUND

Many authors have provided their own insights (Butun, Ismail, Morgera, & Sankar, 2014) and suggested different ways to detect an intrusion in the sensor network. The details of the same are explained in this section.

In (mehmood, et al., 2018) the network is divided into clusters and each cluster is headed by the cluster head. The base station has a blank knowledge database installed. The cluster heads are provided with inference engines. The heads monitor node-related events and data transmission in the respective clusters. The data monitored is considered as events and the same is transmitted to the base station. This

DOI: 10.4018/978-1-7998-3479-3.ch036

received data is analysed to detect suspicious activity. Routine and redundant events are eliminated and same is notified to the respective cluster heads. The threats detected in the received message are notified to the cluster heads. The same is broadcasted to the other cluster members. Any unknown event is analysed and alerted to the cluster heads to avoid transmission. The cluster head is rotated among the members of the cluster for effective and energy consumption of the nodes.

Three stages are proposed by the authors in (Silva, et al., 2005) to detect an intrusion in the network. Filtered messages are collected by the promiscuous node in data acquisition phase. Using this approach the energy is saved. The array storage structure is used to store the discarded messages. In the rule application phase the rules are applied to stored data in array. If the tested data again results in failure, the data is discarded. The approach reduces detection latency. Failure alert is raised if the stored data does not obey the stated rules. An intrusion detection alert is raised if the failure alert is raised beyond a threshold. The failure history is maintained by the respective node. The previous and the present cumulative failure value are combined to obtain new cumulative value. The approach follows deviation failure.

The authors have suggested a security scheme to detect intrusion in (Onat & Miri, 2005). Node impersonation and its effect are addressed in this work. Node impersonation is where the intruder steals the identification of the legitimate node. It provides itself with the same identity credentials. It gains the trust of the other surrounding nodes and gets into depleting the resources or provides false alarms. In the proposed work data and control packets are directional. Tree based forwarding structure is used to route the packets from one end to another. The anomaly report is generated by the neighbours. If more than one node declares illegitimate activities w.r.t its neighbour, the suspicious node is concluded to be guilty.

A Light weight scheme is proposed by the authors in (Ioannis, Krontiris, Dimitriou, & Freiling, 2007) to detect the illegitimate node in the network. The system is focussed to detect blackhole and selective forwarding attacks based on specification based detection. The watchdogs are into cooperative decision making process where the selective forwarding attack is detected to take appropriate actions. The system is into gaining the information from the neighbouring nodes to ward off the guilty node. The nodes are into auditing and aid in building intrusion treat level. This report is shared to bring in collective report. To achieve the success in detection, four conceptual models are utilized. Local packet monitoring, local detection engine, cooperative detection engine and local response are building and the work is divided among the nodes of the network to increase success rate of detecting intruder.

A game theoretic approach is suggested by authors in (Agah, Das, S. K., & Asadi., 2004) to detect intrusion in the network. The authors have suggested three different approaches to detect the compromised node in the network. In the first approach, attack-defence scenario is created. The sensor nodes and the attacker are in non-cooperative mode. The game achieves Nash equilibrium for the attacker and the Intrusion Detection system. This resultant provides an advantage to the Intrusion detection system. Second scheme Markov decision process is used. The system is into learning mechanism. The Intrusion Detection system observes the system and tries to understand the behaviour of the intruder. Using this analogy it tries to analyse which node requires protection. Third scheme uses intuitive method. The node that has the highest traffic load is given the higher priority of protection.

A hierarchical trust based scheme is proposed by the authors in (Chen, Guo, Bao, & Cho, 2014) to detect intrusion in the network. The nodes are evaluated for their honesty, energy and cooperativeness. Sensor node level and cluster head level trust is maintained in the system. The sensor nodes observe other nodes of the cluster for their legitimacy. The cluster head maintains the trust level of other cluster heads in its vicinity. The trust evaluation based on either direct or indirect methodology. The direct evaluation is evaluated for their activity when they communicate with other nodes of the network. The trust level generated by the nodes of the cluster is exchanged by each other. The trust level related to other cluster

heads is forwarded to the base station. The quality of trust and social trust is considered to derive the trust derivative. Stochastic Petri nets are utilized to derive the analytical model.

Local alert modules are used in the study (Krontiris, Benenson, Giannetsos, Freiling, & Dimitriou, 2009) to enable cooperative intrusion detection procedure. The sensors collaborate to identify the malicious node in the network. Different attacking scenarios with identifying the guilty node are discussed in the work. Five stages are used to identify the illegitimate source. One-way key chain containing an ordered list of cryptographic keys is generated using the hash function. These keys are embedded in the nodes before deployment. The nodes send and receive messages from the alert region. This phase is used to authenticate the communicating party. The nodes are to broadcast their key in publish key phase. The other receiving nodes can verify the legitimacy by checking to regenerate the previous key value. In exposing attacker phase the nodes apply a count value for the suspected list and concludes the node as illegitimate based on collaborative data. The actual guilty node is identified in external ring reinforcement stage.

An intrusion detection system is suggested in (Onat & Miri, 2005) for MANETS. The neighbouring nodes in the network alert each other for anomalies. Nodes are programmed to share unexpected behaviour of their neighbours. The detection algorithm uses packet count based sliding window methodology. Last n packets are used to calculate the statistics to make a comparison with the arriving packets. Arriving time and receiver power is recorder. The neighbour has to confirm the statistics of its neighbouring node. The max and min values are benchmarked to the receiver power. An alert is raised if the receiver power crosses either the max or min receiver power. Comparison threshold rate is calculated and if it crosses a certain boundary, an alarm is triggered.

Two factor user authentication is suggested by the authors in (Das & Lal, 2009) to bring in efficiency. The methodology suggests registering the user with the password followed with query phase. The user utilizes a smart card that aids in authenticating him followed by the login procedure.

The authors in (Rajasegarar, Sutharshan, Leckie, Palaniswami, & Bezdek, 2006) has suggested cluster-based anomaly detection algorithm. Euclidean distance is used as dissimilarity measure between pairs of data. Fixed width clustering algorithm is used in the study. After the formation of clusters outliers are classified. The system uses centralized approach. At certain intervals (time window measurement) the sensors send its data to the gateway. The gateway combines its data set with other received set. The cluster algorithm is implemented on this data. The algorithm used minimizes the communication overhead. K nearest neighbour algorithm is used to detect intrusion. The cluster is identified as anomalous if the average cluster distance is more than one standard deviation of inter-cluster distance. The authors have implemented their algorithm on the sensor data collected from great duck island project.

Inside attacker is detection in this work (Fang, Cheng, & Chen, 2007). Four phases make up the algorithm. In the information collection phase the neighbours observe the other nodes in their vicinity and provide the observation result. Q -component attribute vectors are used to provide the result which is broadcasted to its neighbourhood. Packet sending rate, dropping rate, forwarding delay time and sensor readings are measured. False information filtering phase is where the false information obtained by the neighbourhood is filtered. Trust-based false information filtering protocol is used to filter the false information. Trust value is generated by direct observation. If the value is closer to 1 the node has higher probability to be considered as normal sensor. The trust is calculated according to the degree of deviation from the neighbourhood activities. Outlier detection is the third phase of the algorithm. Distance between the sensor to the centre of set data is computed. The mahalanobis distance is calculated. Using the distance variant the suspicious neighbours are listed. To conclude with the decision majority vote is considered.

Intrusion detection in adhoc system is suggested in (Heiderscheit & D., February 25, 2014). The detection of the intrusion is suggested based on messages received from the source. The location details of the received message are compared to derive the authenticity of the source. If any intrusion is happening the same is portrayed in the extracted location detail information. The work also focuses on the communication of the respective node with various sources. The position extracted is shared with other source. The source communicates with an identifier as the resultant of the communication. If the communicated source will be inconsistent if the shared information is unknown to it.

An intelligent intrusion detection methodology is suggested (Thamilarasu & Geethapriya, 2016) to improve security in wireless body network. A multi-objective genetic algorithm is used to detect the attack. Both mobile and wireless sensor nodes are used in the work. Jamming attacks, random jamming attack and selective forwarding attacks are considered in the work. Total number of detections was measured including positive and negative in the system. Three core metrics are used- detection rate, false positive rate and total energy usage. The metrics is used to address the trade-offs between security and resource overhead. The generic algorithm methodologies are used in the proposed work. This includes selection, recombination, mutation and replacement. An optimal combinational of the feature sets is suggested in the training phase.

Beacon frames emitted by unauthorized devices are monitored in (Rathi, M., Jain, Merchant, & Lin., 2014). The edge devices actively participate sending the probe request to trigger response from the unauthorized devices. The received message is verified against the data maintained by intrusion detection system. The Media access control (MAC) address is verified followed by verification of Service Set Identifier (SSID). The system is capable of monitoring packet information. The edge devices record information and transmit the same to the database for further analysis. After finding the malicious edge, the port is switched to preserve security in the network.

Two intrusion detectors are used in the work (Young-Han, et al., 2014). The first detector transmits the classified packet to the second detector. First network processor and packet header intrusion condition are used to classify the packets. The processor does high speed processing using the micro node. The second detector aims to make detection of the packet payload. Using the intrusion detection outcome intrusive packet and non-intrusive packets are collected for future analysis.

A hybrid intrusion detection model is suggested in (Maleh, Ezzati, Qasmaoui, & Mbida., 2015). The system aims in providing high detection rate and low false positive rate. The work is based on support vector machines (SVM). A set of attacks is considered in the work. The system evaluates the malicious behaviour of the target based on the fixed rule of signatures. Cluster-based approach is adopted to increase the lifespan of the nodes. In the work the cluster head behaves as local base station sensor. Two kinds of monitoring agents are considered in the work. The local agent monitors the data transmitted and accepted by the nodes. The signature database is gradually built by observing and using anomaly detection methodology. This information is shared between the nodes of the network. The global agent is used to monitor the communication between the neighbouring nodes.

Low complexity and energy-aware methodology is suggested in (Sudip, Krishna, & Abraham., 2011). The work is based on sampling. The system is designed where the nodes are unaware of their neighbours. The proposal is based on the assumption that the system has a certain budget configured. The nodes are into continuous sampling stage at the minimum budget. The sampling rate is increased if the detection of the malicious activity is more. Once the detection analysed minimizes the sampling rate is reduced.

PROPOSED WORK

Table 1. Notations Used in the Study

Notations used in the work	Description
N	Network under consideration
BS	Base station
N_i	Node i of the network
S_{in}	Set of inferences
Hello	Hello message
Ack	Acknowledgement
r	Transmission radius
CH_i	Cluster head of cluster i
M_A	Mobile agent
H_A	Hash code generated by the cluster head and the mobile agent
A_i	Assistance node of cluster i
N_{id}	Node identification
L_i	Location information of i node
H_i	Hash code generated by i node of the network

Assumptions Made in the Study

The following are the assumptions made in the study-

- The network contains three types of nodes namely cluster head/cluster member, mobile agent and assistance nodes.
- The base station is the most trustable node in the network. It is the responsibility of the base station to provide a set of behaviours assumed to be normal. These sets of behaviour are fed into the assistance nodes before their deployment.
- The nodes deployed in the network are liable to many kinds of attacks. The nodes are free from the attacks till they form clusters.
- The assistance node is the only watchdog that keeps a watch on other nodes of the network. It that does not communicate with other nodes of the network except the base station.

Working of the System

3.2.1 Nodes Used in the Work

- *Static nodes* – The system has three kinds of static nodes.
 - *Cluster members* -These nodes are responsible to sense the environment, process them and transmit the readings.

- *Cluster heads* – These nodes are responsible to aggregate the readings it receives from the cluster members, remove the duplicates and transmit the resultant to the next available hop.
- *Assistance nodes* – These nodes are the reservoirs of reference data and also act as watch dogs.
- *Mobile nodes* – These nodes are used in the system to provide cross-verification status of the suspicious node.

3.2.2 Deployment of the Nodes and Formation of the Clusters

The nodes are deployed randomly in the form of group. Every group will have an assistance node. After the deployment, the nodes broadcast each other with a hello message and its identification. In notation (1) the node N_i is broadcasting hello message and node identification of itself to the network N .

$$N_i \rightarrow \text{Hello}, N_{id} : N \quad (1)$$

The nodes that are within the transmission radius r will be able to hear the message. The nodes that hear the hello message send the acknowledgement message to the transmitter. In the notation (2) N_j is transmitting the acknowledgement message ack and node identification N_{id} to the node N_i .

$$N_j \rightarrow \text{Ack}, N_{id} : N_i \quad (2)$$

The nodes which are able to transmit and receive message will form a cluster. The cluster head is randomly elected. The assistance node is not elected as the cluster head. It is valued as a library that provides and stores set of observations, inferences and behaviours. It is the responsibility of the cluster head to collect the data, eliminate the duplicates and transmit the resultant message to the next available hop or base station. The cluster head is selected based on the trust value provided by the assistance node.

3.2.3 Role of the Assistance Node

The base station would have embedded set of behaviours assumed to be normal in the assistance nodes. The assistance node provides its trust level only at time of electing the new cluster head. The assistance node also behaves as a watchdog and aids in maintaining the trust level of the cluster members. The assistance nodes watch the behaviour of the nodes in the cluster, and make a comparison to the stored behaviour set. Based on this set it concludes a node is legitimate or illegitimate. If it obtains any suspicion on any of the nodes, it notifies the base station. A hash code is generated by using the location information of itself and the node identification. Let H_i be the hash code generated using the location information L_i and node identification N_{id} . In the equation (3) the node N_i is generating the hash code using the available information.

$$N_i : (L_i, N_{id}) \rightarrow H_i \quad (3)$$

The symmetric key is used to encrypt the notification and the hash code is attached to the message before transmission. In equation (4) the node N_i is transmitting the encrypted message E_i suffixed with hash code H_i to the base station BS.

$$N_i \rightarrow E_i \parallel H_i : BS \quad (4)$$

3.2.4 Role of the Mobile Agent

The base station will be able to obtain a reliable message using this approach. On receiving the message from the assistance node the base station alerts the mobile agent. The mobile agent is sent the location derived from the hash code. In the equation (5) the base station BS is alerting the mobile agent M_A with the location information L_i to confirm the illegitimacy of the suspicious node.

$$BS \rightarrow L_i : M_A \quad (5)$$

On receiving the message the mobile agent goes to that location, for some time interval makes the observation of the suspicious node. After a certain interval, the result is transmitted. It also uses its location to derive the hash code and the symmetric key stored inside it to encrypt the transmission data. The hash code is suffixed to the encrypted data and transmitted. In equation (6) the mobile agent M_A is transmitting the encrypted data E_M suffixed with the hash code H_M and time interval T_A to the base station.

$$M_A \rightarrow E_M \parallel T_A \parallel H_M : BS \quad (6)$$

After confirmation, the base station notifies all the nodes of the cluster and the cluster member is isolated from the group.

ANALYSIS OF THE WORK

The proposed work is designed to trace the guilty node and take appropriate action. The nodes in the network are deployed to sense the environment and provide the readings to base station. The cluster is headed by the cluster head that collects the readings from its cluster members and forwards the same to the next available hop. Every cluster has an assistance node that behaves like a watch dog. This node keeps an eye on all the cluster members. Any unusual activity is notified to the base station. The base station notifies the mobile agent for evaluation of the situation. The mobile agent after continuous observation for certain interval confirms the mobile agent with its message.

Assumption That the Cluster Head is Compromised

The cluster head is responsible to aggregate the sensed data transmitted by its cluster members, eliminate the duplicate ones and transmit the resultant to the next available hop. If the cluster head is assumed to be compromised, it will behave unusually. The compromised node will only concentrate to waste the resources and use the collected data inappropriately. It will not be able to do the following –

- It will not be able to communicate with its cluster members w.r.t time
- It will not actively participate in electing the new cluster head
- It will be randomly sending messages to its neighbouring nodes.
- It will not remove the duplicate messages before transmission.

After certain observation, the assistance node notifies the base station. The mobile agent is sent by the base station for re-evaluation. On receiving the message from the mobile agent the base station notifies the suspicious node and the node is isolated from the cluster. A new cluster head is elected.

Assumption That the Cluster Member is Compromised

The cluster member is responsible to sense the environment, process them and transmit the processed data to the cluster head. If the cluster member is compromised it will not be performing its jobs appropriately –

- It will not actively participate in electing the new cluster head
- It will not be able to communicate with its cluster head w.r.t time
- It will be randomly sending messages to its neighbouring nodes.

The assistance node and mobile agent confirms the status. On positive response the base station notifies the cluster of its activities. The guilty node is isolated from the cluster.

Assumption That the Assistance Node is Compromised

The assistance node provides its trust level only at time of electing the new cluster head. It will not communicate with any other node in the network. Hence in most of the cases it is assumed to be reliable. If in case it is compromised the node it will behave inappropriately-

- It will not be able to generate the hash code using its location and node identification.
- It will involve itself in depleting the energy resource by communicating with other nodes of the network
- It will not be able to provide trust level values during electing the new cluster head.

Other nodes of the cluster if receives any message from the assistance node other than the trust level during election of the cluster head notifies the base station. The base station communicates with the mobile agent and assigns the same as the assistance node.

Assumption that the Mobile Agent is Compromised

The mobile agent is used in the proposed work only to evaluate the node for its guiltiness. The mobile agent can behave inappropriately if compromised-

- It will not follow the orders given by the base station. If the orders are not followed by the mobile agent, the base station after not receiving the message from the same concludes it to be inactive or illegitimate. In either of the cases it messages the other mobile agent to perform the job.
- It will not be able to generate the hash code using its location.
- The location of the mobile agent and the location provided by the assistance node have to be within the radius r . If appropriate message is not obtained, the base station terminates the mobile agent and reassigns the job to a new mobile agent.
- It will involve in transmitting messages to other nodes of the network. Other nodes of the network inform the situation to the base station. The base station after receiving a certain count of messages, broadcast the network of the mobile agent.

SIMULATION

The work is simulated using NS2. Table 2 provides the settings used to simulate the work. The work is simulated in 200m*200m dimension. Around 64 static nodes are utilised to collect data from the environment. The nodes are deployed randomly in the environment. 8 clusters are formed. 3 mobile agents are utilised in the work. Total simulation time considered to draw conclusions of the work is 60s.

Table 2. Representation of the setting used in the simulation

settings used for simulation	Description
Node dimension	200m * 200m
Number of nodes deployed	64
Distribution of nodes	Random
Number of clusters	8
Number of mobile agents	3
Simulation time	60 s
Length of encryption key	164 bits
Length of generated hash code	64 bits
Length of broadcast message from base station	256 bits
Length of message by mobile agent and assistance node	284 bits
Length of trust level information provided by the assistance node	24 bits
Length of node identification	16 bits

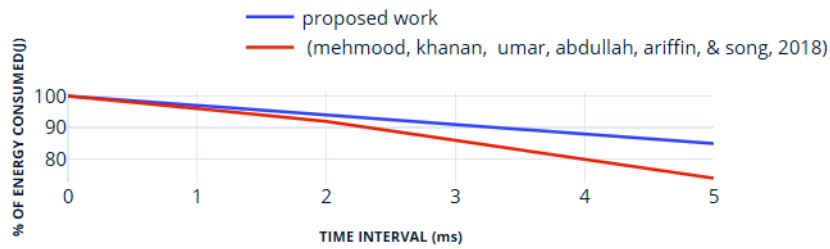
Reliability

In (mehmood, et al., 2018) the network is divided into clusters and each cluster is headed by the cluster head. The base station has a blank knowledge database installed. The cluster heads are provided with inference engines. The heads monitor node-related events and data transmission in the respective clusters. The data monitored is considered as events and the same is transmitted to the base station. This received data is analysed to detect suspicious activity. Routine and redundant events are eliminated and same is notified to the respective cluster heads. The threats detected in the received message are notified to the cluster heads. The same is broadcasted to the other cluster members. Any unknown event is analysed and alerted to the cluster heads to avoid transmission. The cluster head is rotated among the members of the cluster for effective and energy consumption of the nodes.

The proposed work improves reliability compared to (mehmood, et al., 2018). The proposed work considers all the nodes liable to get attacked and hence all the nodes deployed are under supervision. The nodes under the suspicion list is notified earlier and the approach increases reliability.

In the proposed work the assistance node is deployed which does not communicate with other nodes of the network. It only provides the trust level value during electing a new cluster head. Hence to a greater extent it can be considered reliable. A node is concluded as guilty only when it displays certain characteristics and the evaluation is done by the assistance node and the mobile agent. Hence compared to the previous work (mehmood, et al., 2018), the proposed work is 7.07% more reliable. The same is represented in the Figure 1. More the reliability of the network, the security (Lou & Kwon, 2006) (Islam, Shen, & Wang., 2012) also increases.

Figure 1. Comparison of the previous work with the proposed work w.r.t reliability of data received



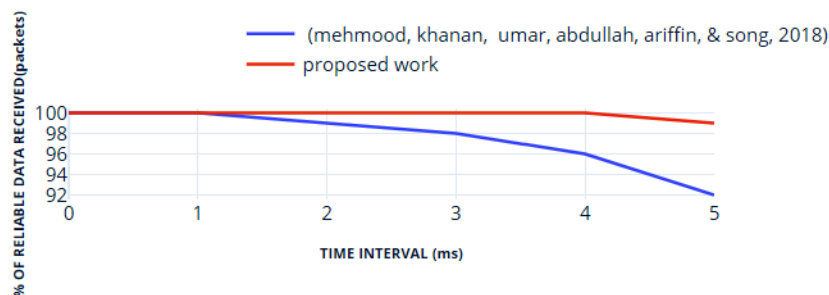
Energy Consumption

Energy is one of the resources which are scarce in the sensors. As these nodes cannot be recharged, caution has to be taken to provide energy conserving algorithms. The proposed work takes the following measures to conserve energy-

- Has adopts Cluster-based routing.
- Duplicate data is eliminated by the cluster head.
- Has adopted Multi-hop routing.
- Only assistance nodes have to behave as watchdog and provide information to the base station. The previous work (mehmood, et al., 2018) makes the cluster head responsible of the notification of the malicious node.
- The cross-verification of messages and evaluation is done more systematically.

Compared to the previous work (mehmood, et al., 2018) the proposed work conserves 12.9% of energy. The comparison w.r.t energy is provided in the Figure 2.

Figure 2. Comparison of previous work to the proposed work w.r.t energy consumption



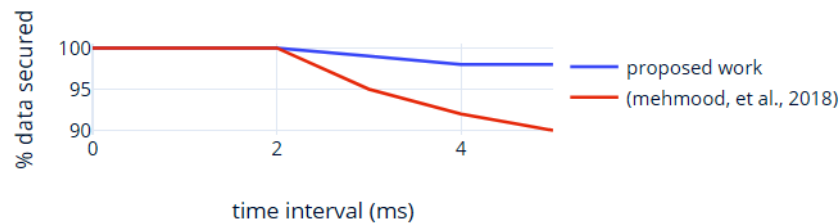
Security

In (mehmood, et al., 2018) the network is divided into clusters and each cluster is headed by the cluster head. The base station has a blank knowledge database installed. The cluster heads are provided with inference engines. The heads monitor node-related events and data transmission in the respective clusters.

The data monitored is considered as events and the same is transmitted to the base station. This received data is analysed to detect suspicious activity. Routine and redundant events are eliminated and same is notified to the respective cluster heads. The threats detected in the received message are notified to the cluster heads. The same is broadcasted to the other cluster members. Any unknown event is analysed and alerted to the cluster heads to avoid transmission. The cluster head is rotated among the members of the cluster for effective and energy consumption of the nodes.

Comparing with the (mehmood, et al., 2018) the proposed work does not rely on cluster head for monitoring information. The assistance nodes in every cluster provide the regular readings of the monitored activities of the nodes of the cluster. If any suspicion they notify the base station. The base station conveys the same to the mobile agent to cross-verify the notification. On affirmation from the mobile agent, the further action is taken. Comparing to the (mehmood, et al., 2018) the proposed work is more reliable and hence the security level is better by 3.025%. The same is represented in the Figure 3.

Figure 3. Illustration of comparative analysis of data secured in both work



FUTURE RESEARCH DIRECTIONS

The proposed work aims in detecting the malicious node in vicinity, notifying the respective and isolating the same on confirmation. Any malicious activity is detected in the proposed work. The assistance node and mobile node aim in detecting and notifying the base station of the activity of the guilty node. The assistance node notifies its suspicion while the mobile node is used to evaluate the notification. The work can be enhanced by-

- If the cluster happens to be located nearby the base station, the energy consumption will be low on transmission. For long distance located nodes, minimizing the energy consumption made by the assistance node for transmitting the notification to the base station has to be considered.

CONCLUSION

Sensor network are the assembly of tiny devices of homogenous or heterogeneous nature. These nodes are capable of sensing the environment, processing the data and transmitting the same. This unsupervised network is always liable to get attacked. Hence some security measures have to be adopted. In addition attacker detection algorithms aid in better reliability and conservation of resources.

The proposed work aims at bringing reliability and conserving energy of the nodes. The system consists of two types of nodes deployed- Static and dynamic. Static nodes except the assistance node senses

the environment, process it and transmits the same. Cluster-based routing is considered in the work. The cluster head aggregates all the data, eliminates the duplicate ones and transmits to next available hop. The assistance node acts as a reservoir of set of behaviours termed as normal for the node. This node watches the other nodes of the cluster and creates a trust level for each. The procedure aids in choosing reliable cluster head. The assistance node provides a report of any malicious activity in its cluster. This report is verified by the base station and mobile agent is summoned for re-evaluation. On confirmation, the node is isolated. All the deployed nodes are tested for their legitimacy in different ways and by different nodes. Hence the reliability is increased by 7.07%, security is increased by 3.025% and energy is conserved by 12.9% compared to the previous work.

REFERENCES

- Agah, Das, A., S. K., B. K., & Asadi., M. (2004). Intrusion detection in sensor networks: A non-cooperative game approach. In *Third IEEE International Symposium on Network Computing and Applications* (pp. 343-346). Cambridge, MA: IEEE.
- Akyildiz, F. I., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: A survey. *Computer Networks*, 38(4), 393–422. doi:10.1016/S1389-1286(01)00302-4
- Arora, A., Dutta, P., Bapat, S., Kulathumani, V., Zhang, H., Naik, V., Mittal, V., Cao, H., Demirbas, M., Gouda, M., Choi, Y., Herman, T., Kulkarni, S., Arumugam, U., Nesterenko, M., Vora, A., & Miyashita, M. (2004). A line in the sand: A wireless sensor network for target detection, classification, and tracking. *Computer Networks*, 46(5), 605–634. doi:10.1016/j.comnet.2004.06.007
- Butun, I., Morgera, S. D., & Sankar, R. (2014). A survey of intrusion detection systems in wireless sensor networks. *IEEE Communications Surveys and Tutorials*, 16(1), 266–282. doi:10.1109/SURV.2013.050113.00191
- Chen, R., Guo, J., Bao, F., & Cho, J. (2014). Trust management in mobile ad hoc networks for bias minimization and application performance maximization. *Ad Hoc Networks*, 19, 59–74. doi:10.1016/j.adhoc.2014.02.005
- Das, & Lal, M. (2009). Two-factor user authentication in wireless sensor networks. *IEEE Transactions on Wireless Communications*, 8(3), 1086-1090.
- Djamel, D., Khelladi, L., & Badache, N. (2005). Security issues of mobile ad hoc and sensor networks. *IEEE Communications Surveys and Tutorials*, 7(4), 2–28. doi:10.1109/COMST.2005.1593277
- Erol-Kantarci, M., & Mouftah, H. T. (2011). Wireless sensor networks for cost-efficient residential energy management in the smart grid. *IEEE Transactions on Smart Grid*, 2(2), 314–325. doi:10.1109/TSG.2011.2114678
- Fang, L., Cheng, X., & Chen, D. (2007). Insider attacker detection in wireless sensor networks. In *26th IEEE International Conference on Computer Communications* (pp. 1937-1945). Barcelona, Spain: IEEE.
- Heiderscheit, & D., D. (February 25, 2014). *Patent No. 8,660,526*. U.S. Patent Office.
- Ioannis, K., Dimitriou, T., & Freiling, F. C. (2007). Towards intrusion detection in wireless sensor networks. *13th European wireless conference*, 1-10.

Islam, K., Shen, W., & Wang, X. (2012). Wireless sensor network reliability and security in factory automation: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(6), 1243–1256. doi:10.1109/TSMCC.2012.2205680 PMID:22438515

Kim, S. S., Culler, D., Demmel, J., Fenves, G., Glaser, S., & Turon, M. (2007). Health monitoring of civil infrastructures using wireless sensor networks. In *6th international conference on Information processing in sensor networks* (pp. 254-263). Cambridge, MA: ACM.

Krontiris, I., Benenson, Z., Giannetsos, T., Freiling, F. C., & Dimitriou, T. (2009). Cooperative intrusion detection in wireless sensor networks. In *European Conference on Wireless Sensor Networks* (pp. 263-278). Springer. 10.1007/978-3-642-00224-3_17

Lee, H. S., Lee, S., Song, H., & Lee, H. S. (2009). Wireless sensor network design for tactical military applications: Remote large-scale environments. In *IEEE Military communications conference (MILCOM)* (pp. 1-7). Boston, MA: IEEE.

Lou, W., & Kwon, Y. (2006). H-SPREAD: A hybrid multipath scheme for secure and reliable data collection in wireless sensor networks. *IEEE Transactions on Vehicular Technology*, 55(4), 1320–1330. doi:10.1109/TVT.2006.877707

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., & Anderson, J. (2002). Wireless sensor networks for habitat monitoring. In *1st ACM international workshop on Wireless sensor networks and applications* (pp. 88-97). Atlanta, GA: ACM.

Maleh, Y., Ezzati, A., Qasmaoui, Y., & Mbida, M. (2015). A global hybrid intrusion detection system for wireless sensor networks. *5th International Symposium on Frontiers in Ambient and Mobile Systems*, 52, 1047-1052. 10.1016/j.procs.2015.05.108

Mehmood, A., Khanan, A., Umar, M. M., Abdullah, S., Ariffin, K. A., & Song, H. (2018). Secure knowledge and cluster-based intrusion detection mechanism for smart wireless sensor network. *IEEE Access*, 5688-5694.

Onat, I., & Miri, A. (2005). An intrusion detection system for wireless sensor networks. *IEEE International Conference on Wireless And Mobile Computing, Networking And Communications*, 3, 253-259. 10.1109/WIMOB.2005.1512911

Rajasegarar, S., Leckie, C., Palaniswami, M., & Bezdek, J. C. (2006). Distributed anomaly detection in wireless sensor networks. In *10th IEEE Singapore international conference on communication systems* (pp. 1-5). Singapore: IEEE.

Rathi, M. M., Jain, V. K., Merchant, S. T., & Lin., V. C. (2014). *Patent No. 8,707,432*. U.S. Patent Office.

Silva, d., R., A. P., Martins, M. H., Rocha, B. P., Loureiro, A. A., Ruiz, L. B., & Wong., H. C. (2005). Decentralized intrusion detection in wireless sensor networks. In *1st ACM international workshop on Quality of service & security in wireless and mobile networks* (pp. 16-23). Montreal, Canada: ACM.

Sudip, M., Krishna, P. V., & Abraham, K. I. (2011). A simple learning automata-based solution for intrusion detection in wireless sensor networks. *Wireless Communications and Mobile Computing*, 11(3), 426–441. doi:10.1002/wcm.946

Sun, B., Osborne, L., Xiao, Y., & Guizani, S. (2007). Intrusion detection techniques in mobile ad hoc and wireless sensor networks. *IEEE Wireless Communications*, 14(5), 56–63. doi:10.1109/MWC.2007.4396943

Thamilarasu & Geethapriya. (2016). iDetect: an intelligent intrusion detection system for wireless body area networks. *International Journal of Security and Networks*, 11(1-2), 82-93.

Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. *Computer Networks*, 52(12), 2292–2330. doi:10.1016/j.comnet.2008.04.002

Young-Han, C., Kim, D., Lee, S.-R., Lee, M., Bae, B.-C., & Park, S. (2014). *Patent No. 8,732,833*. U.S. Patent Office.

KEY TERMS AND DEFINITIONS

Assistance Node: These are nodes deployed with other nodes in the network to aid the other nodes in their doings.

Cluster-Based Routing: The nodes in the network are grouped together in the form of clusters. The nodes transmit their readings to the group head which later transmits the aggregated data to the next available hop.

Intrusion Detection Scheme: The methodology adopted aids in detecting the compromised nodes in the network.

Knowledge-Based Anomaly IDS: The nodes after certain amount of observation of the network are provided with certain observations that aid them in locating illegitimate nodes.

Mobile Agent: Mobile agents are the set of dynamic nodes capable to moving from one end of the network to another and doing their regular routines similar to static nodes of their calibre.

Smart Environment: Smart environment is an assembly of sensors, actuators and many computational elements providing services for betterment of human life.

Wireless Sensor Network: Wireless sensor network (WSN) refers to a group of spatially dispersed and dedicated sensors deployed to monitor/track any object of interest.

Improved Cross-Layer Detection and Prevention of Sinkhole Attack in WSN

Ambika N.

 <https://orcid.org/0000-0003-4452-5514>

Department of Computer Application, SSMRV College, Bangalore, India

INTRODUCTION

Wireless sensor network (Chen, Makki, Yen, & Pissinou, 2009) (I.Akyildiz, Su, & Sankarasubramanian, 2002) is an assembly of tiny devices which self-configure and communicate in the unsupervised environment. These devices are used in many applications (Xu, 2002). Military surveillance (Lee, Lee, Song, & Lee, 2009), agriculture monitoring (Baggio, 2005), habitat monitoring (Polastre, Szewczyk, Mainwaring, Culler, & Anderson, 2004), smart homes (N.K.Suryadevara & Mukhopadhyay, 2012), hospital surveillance (López, Custodio, & Moreno, 2010) are some of the applications where these devices play a major role in accumulating data without human intervention.

Transmitting confidential information is one of the responsibilities of these devices. While doing so, these devices are liable to get compromised due to their limitations. As these devices are not monitored, the illegitimate nodes and their activities will remain undiscovered. Hence to bring the guilty into light better security measures have to be accommodated. The methodology adopted has to minimize the occurrence of the same by alerting the legitimate nodes of their activity.

Sinkhole attack (Ngai, Liu, & Lyu, 2006) (Rehman, Rehman, & Raheem, 2019) is one such attack where the compromised nodes advertise themselves. They broadcast false hop count to the normal nodes in their vicinity. The attack is launched by the illegitimate node advertising itself as the most favourable path to the base station. Their activity not only attracts the traffic towards itself but also utilizes the network resources. The guilty node after obtaining the trust of its surrounding nodes attracts all the traffic and is liable to launch different kinds of attacks. In the proposed work the solution is provided to minimize sinkhole attack in the network. The proposed work is improving cross-layer design suggested in (S & S, 2017), where Media Access control layer (MAC) and network layer is provided with algorithms to perform. The proposed work is divided into five sections. A detailed literature survey is provided in section 2. The proposed work is explained with detail analysis in section 3. Future research directions are suggested in section 4. The work is concluded in section 5.

BACKGROUND

The proposed work provides solution to tackles and minimizes the occurrences of Sinkhole attack in the network. Many authors have provided their suggestions to tackle the same. This section provides a brief history of the same.

In (Ngai, Liu, & Lyu, 2006) network flow graph is utilized to detect sinkhole attack. The selective forwarding activity is considered to make the detection of the attack. The nodes are monitored to find the missing data from the region. A statistical methodology is adopted to accomplish the task. Using the

DOI: 10.4018/978-1-7998-3479-3.ch037

mean of the data sent and amount of data collected using sliding window a measure is calculated. This calculated value is compared against the threshold to measure the suspicious activity in the respective area. The base station gauges the suspected area. After detecting the affected area the base station forwards a request message to the area. The message contains the IDs of all the affected nodes and is made to move from hop to hop. A timestamp is included in the message to avoid replay of the messages. The respective nodes sign the message with its private key. After receiving the message from the nodes, the receiving reply with the base station with its message. The message contains details of IDs, ID of the next hop and estimated cost [hop-count, data rate]. The base station constructs the tree using the information from the received message. The area under sinkhole attack follows a pattern where all the packets move towards the invaded node. Using this methodology the adversary can be detected.

The authors in (Krontiris, Dimitriou, Giannetsos, & Marios, 2007) have constructed a routing tree to tackle sinkhole attack. The system follows distributed architecture. Identical IDS clients are embedded in the nodes deployed in the network. The IDS agent is responsible for network monitoring, intrusion detection and decision making. The client listens to the network traffic, captures the individual packets and examines the same. The packets in vicinity are scrutinized for legitimacy. A specification based approach is followed by the supervising nodes. A deviation from the normal user defined rules is considered in the detection procedure. The system has to provide appropriate rules to detect suspicious behaviour. Two rules are used in the proposed work. In both the rules the sender field is verified to know its identity. The node ID has to be one of the neighbours and not the examining node. A cooperative mutual conclusion approach is followed to conclude the behaviour of the guilty nodes. Each watchdog in the neighbouring list broadcast the alert message and arrives at the conclusion. On affirmation, the nodes involve themselves in regenerating the cryptographic materials and distributing the same among the trusted nodes. The base station is notified of the illegitimate node.

In (S & S, 2017) cross layer approach is adopted. The paper tackles sinkhole attack. LEACH protocol is used in the work. Two assumptions are considered in the work. It is assumed that cluster heads cannot be compromised. The nodes are availed flexibility to vary their transmitting power. Rectangular clustering approach is adopted. The network is divided into rectangular clusters of equal area. The average location of nodes is calculated using Dijkstra's algorithm. The MAC layer sends link information to the network layer. The information is evaluated considering re-transmissions. Using the estimations an optimal routing scheme is provided. Considering Packet-delivery ratio sinkhole attack is evaluated. The work considers only retransmission to locate sinkhole attack.

The authors have proposed a different way to detect sinkhole attack in (Krontiris, Giannetsos, & Dimitriou, 2008). Two scenarios are considered in the work. Mintroute protocol (Rassam, Zainal, Maarof, & Al-Shaboti., 2012) is considered where the routing tree is built considering the link quality estimates. Packet error rate is considered. Periodically an update packet is transmitted to check the packet loss based on the number of packets received by the neighbouring nodes. The data of the same is updated in all the nodes with the respective sender node IDs. The nodes under sinkhole attack persuade the other legitimate nodes to change their route and choose the adversary as its next hop. Once the message is received by the legitimate node, the same checks the received message. If the node is not the one of the legitimate node in its list, it rejects the offer. Second scenario considered in the work is Multihop Link Quality Indicator Protocol (MultiHopLQI) (Gupta, Sharma, Marot, & Becker, 2010). The nodes calculate the link quality based on its own hardware. A beacon message is broadcasted by the nodes. The receiver uses its hardware to calculate LQI. Link quality indicator provides the cost of the corresponding link. The payload is the summation of cost of all links that make the given path. The nodes in this environment check the advertised path cost of the node and makes a comparison with the path cost

with its parent. The comparison resulting in positive outcome aids the node to change its parent. This approach aids in immediate detection of the illegitimate node. The parent-child link LQI is measured and ensured not to exceed 10.

An efficient intrusion detection algorithm is suggested in (Ngai, Liu, & Lyu.M.R, 2007). Two conditions are considered in the work. The network having single and multiple adversaries is discussed. The dropping of packets in multiple adversary environments is discussed in the work. The sensors dispatch the information to the base station using multiple redundant paths. Analyzing the outcome of the normal nodes is detected in their vicinity. The constructed tree is broken down into sub trees. The illegitimate node denying the packet transmission is detected. To deal with tampered or false messages a different approach is proposed in the work. A secret key is communicated along with message authentication code. The hop counts are calculated is made a comparison with number of edges from the given node to the current node. Identifying inconsistency in the counts the illegitimate node is detected.

Two approaches are suggested by the authors to tackle sinkhole attack in (Shafiei, Khonsari, Dera-khshi, & Mousavi, 2014). A centralized approach is adopted to detect suspicious regions in the network. A geostatistical model is utilized to do so. Distributed monitoring approach is adopted. The base station periodically samples the deployment area and collects the residual energy from sampled nodes. Along with the residual energy geostatistical data is also collected. This data is used to measure criticality of the sampled region. Based on the analysis report the base station takes appropriate action. A geostatistical frailty survival model is utilized to detect energy holes. The model aids in establishing the connection between before and after the occurrence of an event. Using distributed monitoring approach details of guilty node is learnt. The aggregate rate of the packets and the energy used to transmit the packet is calculated. The difference in energy against the threshold puts the monitored node into suspicion. The base station on confirmation takes appropriate action.

The authors in (Salehi, Razzaque, Naraei, & Farrokhtala, 2013) have suggested detecting sinkhole attack in the network. The work adopts a methodology to detect a single malicious node and multiple occurrences of the same at the later stage. The algorithm verifies data consistency to find the guilty nodes in the network. Using network flow information the malicious node is detected. A low-overhead algorithm is used to detect the attacked area. The identification algorithm is followed after detecting the malicious node residing group. The nodes identify themselves by providing the Ids of their own and the next-hop node. As flooding procedure is adopted, the most used hop is identified. Using the information, the attacker detected.

In (Tumrongwittayapak & Varakulsiripunth, 2009) the authors have suggested a lightweight and robust algorithm to detect sinkhole attack. The algorithm is based on received signal strength indicator (RSSI) readings. Extra monitor nodes are used to obtain good amount of readings. The distance between the transmitter and receiver is measured based on transmit power with effective propagation loss. Lognormal shadowing model is used to calculate the transmission power. Visual geographic map is calculated by using RSSI value. The monitoring nodes aid in the process. The base station broadcasts hello message to the network. The sensors use the predefined information in the routing table to locate the next hop. Based on the transmission the monitor nodes draw its conclusion based on the received message.

One-way hash chain is used to detect sinkhole attack in (Sharmila & Umamaheswari, 2011). The proposed work uses padding logic for MD5 extension. The message digest of 160 bits length is calculated. This message digest is attached to the original message. The single block of message is sent as input to secure hash algorithm and message digest algorithm. The digest obtained acts as the initial value for next block message. The message after passing through the malicious node is liable to get

modified. This is compared with the digest forwarded by the trustable node. The comparison result aids in locating the guilty node.

A non-cryptographic method was suggested by the authors in (Sheela, Kumar, & Mahadevan, 2011). The proposed work does not require any encryption or decryption mechanism. The mobile agents are used to accomplish the task. The agent is a program segment that aids in self controlling. They move from one end of the network to another transmitting data. The designed system aids the valid node from listening to the cheating information from compromised node. Two algorithms are proposed in the work. The mobile agent carries some dummy data inside. The alteration of dummy data confirms corruption. The least visited neighbour list is jolted. The updates are amended in the cache memory when the agent visits the node. Data routing algorithm is designed using data stored in the mobile agent. The algorithm aids in finding the maximum repeated hop with less weight.

In (Qi, Hong, Xiaohui, & Qiang, 2012) multihopLQI protocol is used. The protocol calculates the link quality and aids in amending the current parent. The system detects the sinkhole attack using two methods. In the first method the received BeaconMsg is checked to have ID of the neighbour. In the second method the path cost of parent/child is checked not to cross 10. The work suggests using encryption/decryption mechanism or creating a blacklist.

A novel intrusion detection system is suggested to detect sinkhole attack in (Dallas, Leckie, & Ramamohanarao, 2007). The proposed work considers guarding base station and the static nodes deployed in the network. Directed flood routing protocol is used in the proposed work. The base station broadcast the route advertisement to the sensors. The broadcast makes the sensors aware of the data collection point. The hop-count is updated in all the nodes. The methodology used aids in detecting the node under sinkhole attack. An alert is created which aids the respective nodes to take appropriate action on time.

A solution is suggested to detect sinkhole and wormhole attack in (Pirzada & McDonald, 2005). Dynamic Source routing protocol (DSR) routing is considered in the work. All the packets are affixed with DSR source route header. The list contains the complete route details the packet has to traverse to reach the destination. The source node broadcast the message with the destination address. The nodes listening to the broadcast, in-turn broadcast the same message affixing its address to the message. The broadcast is done in non-propagating manner to avoid repetition. The reply packet is generated by the respective transmitting node and broadcasted using the list obtained in the received packet. The work uses trust model approach. The model creates trust levels of the respective nodes in communication. The nodes monitor each other to measure accuracy and sincerity of its neighbouring nodes. An integrity check is done when a node receives a packet. The increment/decrement of the trust is based on the outcome of the analysis done by the monitor node.

The authors in (Le Fessant, Papadimitriou, Viana, Sengul, & Palomar, 2012) have suggested some measures to create resilience against the compromised nodes in the network. The proposed work is suggested to bring in solution for selective-forwarding and sinkhole attacks. Three tree-based routing protocols are studied in the proposed work. Hop distance is used to build the routing tree. Periodic routing aids in updating the routing topology regularly. Two resilient protocols are suggested in the work – RESIST-1 and RESIST-0. RESIST-1 protocol aids in preventing the illegitimate node from modifying its advertised distance. RESIST-0 does not allow manipulation of the advertisement at the cost of additional complexity.

A precise solution is suggested in (Stafrace & Antonopoulos, 2010) to detect sinkhole attack in short duration of time. To calculate precision ratio of true detection to the false ones are calculated. False detections are inclusive of false negatives and positives. The proposed work uses AODV protocol to establish route between two nodes.

A trust based intrusion detection system is proposed in (Ghugar, Pradhan, Bhoi, & Sahoo, 2019). Three layers are considered to calculate the trust value. At the physical layer denial of service attack is considered. The nodes are considered to accept malicious packets along with the expected ones. The proposed system assumes that the number of packets generated by the malicious node is always greater than that of the normal node. A pseudo random number uniform distribution pattern is analysed. To detect the malicious activity, energy consumed and number of messages metrics is obtained from the monitoring nodes. The second layer into consideration in the work is MAC layer. Back-off manipulation attack is considered in the MAC layer. To detect malicious node the back-off time parameter and number of successful message transmission parameter are considered. Network layer is the third layer considered in the proposed work. The malicious node in the proposed work is assumed to broadcast fake messages w.r.t hop count. Route reply packet is used to perform so. To detect the malicious activity hop count metrics is utilised.

The Negative selection algorithm is used in (Zhang & Xiao, 2019). Two kinds of infrastructure are used – flat infrastructure and hierarchical infrastructure. The intrusion detection system is used for the hierarchical model. The cluster head communicates with the base station. The nodes deployed are divided into three layers- application layer, immune layer and network layer. The cluster head commences with the immune layer. This layer is responsible to monitor the transmission flow in the cluster members and neighbouring cluster heads. The work uses variable-radius detectors. The negative selection algorithm analyses the distribution. Real-valued spaces are used to find the subspaces. The detector is activated based on the algorithm search. A mature detector is located if it crosses the tolerance.

PROPOSED WORK

Notations Used in the Study

The table 1 contains the description of the notations used in the proposed study.

Table 1. List of notations used in the study

Notations	Description
N	Network under study
N_i	i th node in the network
r	Assumed transmission range to receive communication
Ack	Acknowledgement message
Hello	Hello message
CH_i	Cluster head of i th cluster
U_{id}	Unique identification of the node
BS	Base station
K_i	Public key
S_{bs}	Signal used by the base station
H_i	Hash code
T_i	Sensed data to be transmitted

Assumptions Made in the Study

The following assumptions are made in the proposed work:

- The network is assumed to be free from any kind of attack till the nodes configure and commences the communication with each other.
- The adversary is capable of introducing sinkhole attack into the network. The nodes under this kind of attack will advertise itself as the most favourable path to the base station. After gaining trust of the legitimate nodes, they are capable of launching different kinds of attacks into the network.
- The base station is assumed to be most trustable node in the network. All the other nodes of the network are liable to get compromised.

Deployment of Nodes in the Network

The base station embeds the keying materials into the nodes prior to the deployment. The nodes in the network are deployed randomly. The nodes self-configure and broadcast Hello message. In notation (1) node N_i is broadcasting Hello message to the network N .

$$N_i \rightarrow \text{Hello} : N. \quad (1)$$

The nodes within the transmission range r will be able to receive the transmitted message. They respond with the acknowledgement message. In notation (2) node N_j is acknowledges with the Ack message to node N_i .

$$N_j \rightarrow \text{Ack} : N_i. \quad (2)$$

These nodes form the cluster. A cluster head is chosen randomly in the first round. The head selection is rotated to other nodes in the network to increase the lifespan of the nodes. The cluster head is chosen based on the residual energy in them.

Finding the Route to the Base Station

The base station broadcasts public key to the network. In the equation (3) the base station BS is broadcasting public key K_i to the network N .

$$BS \rightarrow K_i : N. \quad (3)$$

The base station broadcasts Hello message along with the public key to the nearest nodes using a particular signal. The signal is the same used by the sensors of the static nodes to communicate with each other. In equation (4) the base station is broadcasting Hello message and public key K_i to the network using the signal strength S_{bs} .

$$BS \rightarrow S_{bs} (Hello, K_i) : N. \quad (4)$$

The nodes which are able to listen to the signal respond with the acknowledgement message. The nodes encrypt the node identification using the public key and transmit the same to the base station. In the equation (5) the base station BS is broadcasting Hello message and unique identification encrypted using public key K_i to the base station BS.

$$N_i \rightarrow (Ack : K_i (U_{id})) : BS. \quad (5)$$

The base station after receiving the message from the nodes decrypt using the private key and identify the message for its legitimacy. The authenticated nodes broadcast the Hello message to other nodes in the network with the same signal strength. It receives the acknowledgement message along with encrypted node identification message. This is forwarded by the respective nodes to the base station. The base station evaluates them for their legitimacy and broadcasts if it finds any malicious nodes in the network.

In this way all the nodes in the network are authenticated and their location is traced. This procedure is repeated at certain intervals of time and the data is updated. All the nodes maintain the list of legitimate/illegitimate nodes and the path is updated.

Attaching the MAC Message to the Transmitted Message

A hash code is generated using the node identification. Both the sensed message and the hash code are encrypted using the public key. In equation (6) the message T_i and generated hash code H_i is encrypted using the public key K_i and transmitted to the next authenticated hop N_j .

$$N_i \rightarrow K_i (T_i || H_i) : N_j. \quad (6)$$

To ensure security in the network, the hash code is changed for every session. The receiving node forwards the message to the base station. The base station on receiving the message decrypts the message using its private key. The hash code is cross-verified for its legitimacy. The base station ensures that it transmits different public key to the network for every session. The measure increases security of the data in the network. The methodology implements forward secrecy.

Analysis of the Work

The work is suggested to minimize sinkhole attack in the network. The base station authenticates the nodes of the network by evaluating the transmitted encrypted node identification. The signal strength broadcasted gauges the distance between each other. The suggested work aids the node to find appropriate path for transmission. The compromised node can be detected in the following manner.

- *Assuming the cluster head is compromised* – The cluster heads transmits the message to the authenticated nodes of the network. If the head is compromised it will either falter while authenticating or while transmitting the message to the next available hop. The base station will be able to find the malicious node after receiving the message. On affirmation of the illegitimacy, the nodes

will be notified for its activity. The nodes make an entry of the guilty node and choose another node as its cluster head.

- *Assuming the cluster member is compromised* – The cluster members sense the environment and transmit the data by encrypting them using public key. The cluster members encrypted hash code is also attached to the transmitted message. This is evaluated by the base station. The hash code has to change with every session and has to sync with the node identification. The compromised node will not be able to do so and hence the malicious activity can be detected. After receiving notification from the base station, the malicious node is isolated from the cluster.

Simulated Work

The work is simulated using NS2. Various parameters are used in the proposed study. The dimension of the network considered is 500m *500m. The nodes are deployed randomly. 100 nodes are deployed in the network forming 12 clusters. The same is represented in the Table 2.

Table 2. List of parameters used in the proposed work

<i>Parameters</i>	<i>Description</i>
Dimension	500m*500 m
Number of static nodes deployed in the network	100
Distribution of nodes	Random
Number of clusters	12
Length of sensed message	840 bits
Length of the hash code generated	84 bits
Length of private key generated by base station	240 bits
Length of public key generated by the base station	240 bits
Time interval	60s

Energy Consumption

Energy (Fabri.F, Buratti.C, & Verdone.R, 2008) is one of the limited resources in sensors. Hence consuming the same becomes essential. The following methodologies are utilized to increase lifespan of the nodes in the network-

- Multi-hop transmission is adopted- The cluster head transmit their packets to the next available hop and hence large amount of energy is conserved.
- Cluster based transmission (Xiangning, 2007) is adopted- The proposed work adopts cluster based routing. The nodes able to hear each other are grouped together. The cluster members transmit their packets to the group head. The group head transmits the aggregated packets to the next available hop.
- Authentication (Abraham & Ramanatha, 2006) (Butun & Sankar, 2011) is performed for every certain time interval. This is the preliminary stage where the nodes are able to identify each other before the transmission commences. Illegitimate nodes are traced at an earlier stage. The nodes

undergo authentication for certain interval of time and hence the malicious nodes can be detected and isolated.

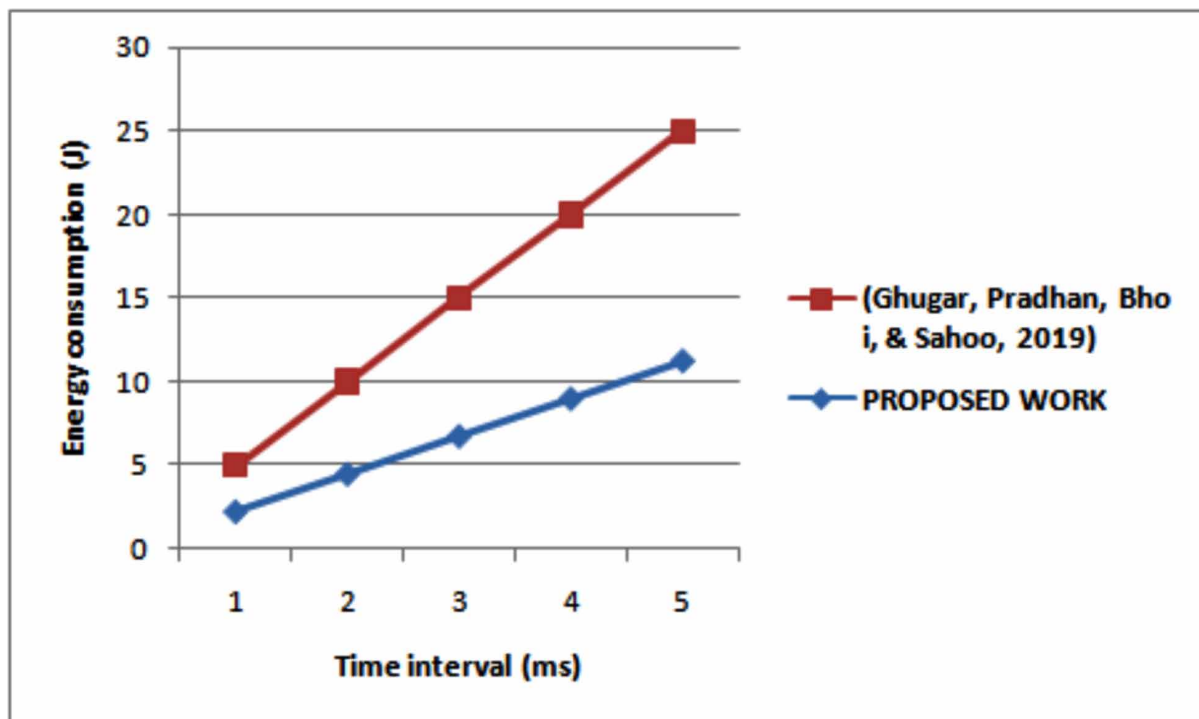
The proposed work is compared to (Ghugar, Pradhan, Bhoi, & Sahoo, 2019). The base station in the proposed work broadcasts the public key which is later used by the nodes to-

- Encrypt hash code
- Encrypt transmitted message

The base station decrypts the encrypted message using its private key. As only the base station will have the decryption key, the communication would remain secure. The base station repeats the broadcasting process for every session. Hence security for the transmitted data is enhanced. Increased security keeps away the illegitimate nodes and reduces unnecessary use of resources.

A trust based intrusion detection system is proposed in (Ghugar, Pradhan, Bhoi, & Sahoo, 2019). Three layers are considered to calculate the trust value. At the physical layer denial of service attack is considered. The nodes are considered to accept malicious packets along with the expected ones. The proposed system assumes that the number of packets generated by the malicious node is always greater than that of the normal node. A pseudo random number uniform distribution pattern is analysed. To detect the malicious activity, energy consumed and number of messages metrics is obtained from the monitoring nodes. The second layer into consideration in the work is MAC layer. Back-off manipulation attack is considered in the MAC layer. To detect malicious node the back-off time parameter and number of successful message transmission parameter are considered. Network layer is the third layer

Figure 1. Comparison of energy consumption in the network



considered in the proposed work. The malicious node in the proposed work is assumed to broadcast fake messages w.r.t hop count. Route reply packet is used to perform so. To detect the malicious activity hop count metrics is utilised. Taking entire network into consideration, the total trust value to be calculated is

Trust value = number of nodes * number of transmissions * length of the message * number of clusters in the network.

Energy consumption at receiving end = (number of nodes * energy consumed by receiving nodes)

Energy consumed at transmitting end = (number of nodes * number of transmissions * energy consumed for transmission)

Total energy consumed = (Energy consumption at receiving end + Energy consumed at transmitting end) * number of clusters

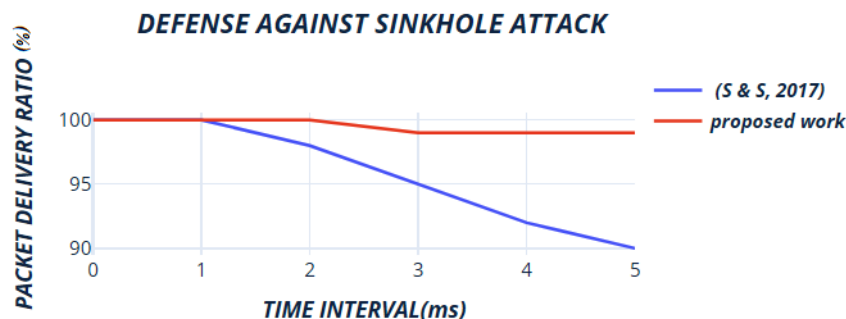
In (Ghugar, Pradhan, Bhoi, & Sahoo, 2019) three layers – physical layer attack, MAC layer attack and network layer attack is considered. The energy consumption to make the nodes secure against these attacks is higher by 18.18% compared to the proposed work. The same is represented in the fig 1.

Security Against Sinkhole Attack

The work proposed by the authors in (S & S, 2017) uses Dijkstra's algorithm to find the location of the nodes. It suffixes the MAC message to authenticate itself. The proposed study is an improvement over (S & S, 2017). Two measures are adopted to safeguard the data against the Sinkhole attack in the proposed work-

- The proposed work uses broadcasting procedure where the base station authenticates the nodes nearer to itself. The legitimate nodes authenticate other nodes of the network within the signal range. Hence all the nodes get authenticated and this procedure is performed for every session. The node provides their authentication by transmitting the node identification in the encrypted form. As the base station has the private key, it decrypts the received message to authenticate them.
- The nodes also suffix the generated and encrypted hash code to the transmitted message. The methodology brings in better security to the transmitted data by enhancing security by 22% with comparison to the work (S & S, 2017). The graphical representation for the same is provided in figure 2.

Figure 2. Comparison of security against sinkhole attack



FUTURE RESEARCH DIRECTIONS

The proposed work uses broadcasting procedure where the base station authenticates the nodes nearer to itself. The legitimate nodes authenticate other nodes of the network within the signal range. Hence all the nodes get authenticated and this procedure is performed for every session. The node provides their authentication by transmitting the node identification in the encrypted form.

The base station broadcasts the public key which is later used by the nodes to encrypt hash code and encrypt transmitted message. The base station decrypts the encrypted message using its private key. As only the base station will have the decryption key, the communication would remain secure. The base station repeats the broadcasting process for every session. Hence security for the transmitted data is enhanced. Increased security keeps away the illegitimate nodes and reduces unnecessary use of resources. To enhance the proposed work –

- The receiving nodes also consume energy to accept the transmitted message. As the number of transmission by the base station is increased to bring in security to the network increases, the nodes also consumes lot of energy. Reducing the energy consumption can be considered in future work.

CONCLUSION

Wireless sensor network are devices used in many applications with varying utilization. These unsupervised devices require adopting some strong mechanism to provide secure communication to data. Sinkhole attack is one such attack where the adversary advertises itself as one of the best hops available to reach the base station. The legitimate nodes unknowingly will share the key credentials and other information.

The proposed work is suggested to bring in better security to the network by detecting the malicious node at an early stage. The base station uses the signal strength to draw in the acknowledgement from the nodes that fall into the inner circle. Using the signal strength aids in measuring distance between the nodes. The legitimate nodes in-turn authenticates the respective nodes using the same signal strength. The procedure aids in authenticating the entire network. The base station broadcasts public key to the network for every session. This key is used to encrypt the transmitted data, encrypting the node identification during authentication procedure and encrypting the hash code. The hash code is changed for every session to enhance security in the network. As the private key only resides with the base station, only the same will be able to decrypt the message. The procedure aids in increasing forward secrecy. It enhances security towards sinkhole attack by 22% and minimizes energy consumption by 18.18% with a comparison to previous work. The methodology increases confidentiality and integrity to data.

REFERENCES

- Abraham, J., & Ramanatha, K. S. (2006). An Efficient Protocol for Authentication and Initial Shared Key Establishment in Clustered Wireless Sensor Networks. *Proc. Third International conference on Wireless and optical Communication networks*, 5-10. 10.1109/WOCN.2006.1666568
- Akyildiz, Su, & Sankarasubramaniam. (2002). A Survey On Sensor Networks. *IEEE Communications*, 102-114.

- Baggio, A. (2005). Wireless sensor networks in precision agriculture. In *ACM Workshop on Real-World Wireless Sensor Networks (REALWSN 2005)* (Vol. 20). Stockholm, Sweden: ACM.
- Butun, I., & Sankar, R. (2011). *Advanced Two Tier User Authentication Scheme for Heterogeneous Wireless Sensor Networks*. In *IEEE Consumer communications and networking conference*. IEEE.
- Chatterjee, K., De, A., & Gupta, D. (2015). A Secure and Efficient Authentication Protocol in Wireless Sensor Network. *Wireless Personal Communications*, 81(1), 17–37. doi:10.1007/11277-014-2115-2
- Chen, X., Makki, K., Yen, K., & Pissinou, N. (2009). Sensor network security: A survey. *IEEE Communications Surveys and Tutorials*, 11(2), 52–73. doi:10.1109/SURV.2009.090205
- Dallas, D., Leckie, C., & Ramamohanarao, K. (2007). Hop-Count Monitoring: Detecting Sinkhole Attacks in Wireless Sensor Networks. In *15th IEEE International Conference on Networks* (pp. 176–181). Adelaide, Australia: IEEE.
- Fabri, F., Buratti, C., & Verdone, R. (2008). A multi-sink multi-hop wireless sensor network over a square region: connectivity and energy consumption issues. *Proc. GLOBECOM Workshop*, 1–6. 10.1109/GLOCOMW.2008.ECP.38
- Ghugar, U., Pradhan, J., Bhoi, S. K., & Sahoo, R. R. (2019). LB-IDS: Securing Wireless Sensor Network Using Protocol Layer Trust-Based Intrusion Detection System. *Journal of Computer Networks and Communications*, 1–14.
- Gupta, A., Sharma, M., Marot, M., & Becker, M. (2010). Hybridlqi: Hybrid multihoplqi for improving asymmetric links in wireless sensor networks. In *Sixth Advanced International Conference on Telecommunications* (pp. 298–305). Barcelona, Spain: IEEE. 10.1109/AICT.2010.58
- Krontiris, I., Dimitriou, T., & Giannetsos, T., & Marios. (2007). Intrusion Detection of Sinkhole Attacks in Wireless Sensor Networks. In *International symposium on algorithms and experiments for sensor systems, wireless networks and distributed robotics* (pp. 150–161). Springer.
- Krontiris, I., Giannetsos, T., & Dimitriou, T. (2008). Launching a sinkhole attack in wireless sensor networks; the intruder side. In *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications* (pp. 526–531). IEEE. 10.1109/WiMob.2008.83
- Le Fessant, F., Papadimitriou, A., Viana, A. C., Sengul, C., & Palomar, E. (2012). A sinkhole resilient protocol for wireless sensor networks: Performance and security analysis. *Computer Communications*, 35(2), 234–248. doi:10.1016/j.comcom.2011.09.005
- Lee, S. H., Lee, S., Song, H., & Lee, H. S. (2009). Wireless sensor network design for tactical military applications: Remote largescale environments. In *Proc. IEEE conference in Military Communications* (pp. 1–7). IEEE. 10.1109/MILCOM.2009.5379900
- López, G., Custodio, V., & Moreno, J. I. (2010). LOBIN: E-textile and wireless-sensor-network-based platform for healthcare monitoring in future hospital environments. *IEEE Transactions on Information Technology in Biomedicine*, 14(6), 1446–1458. doi:10.1109/TITB.2010.2058812 PMID:20643610
- Ngai, E., Liu, J., & Lyu, M. R. (2007). An efficient intrusion detection algorithm against sinkhole attacks in wireless sensor network. *Computer Communications*, 30(11–12), 2353–2364. doi:10.1016/j.comcom.2007.04.025

- Ngai, E. C., Liu, J., & Lyu, M. R. (2006). On the intruder detection for sinkhole attack in wireless sensor networks. In *IEEE International Conference on Communications* (pp. 3383-3389). IEEE.
- Ngai, E. C., Liu, J., & Lyu, M. R. (2006). On the intruder detection for sinkhole attack in wireless sensor networks. In *IEEE International Conference on Communications* (pp. Vol. 8, pp. 3383-3389). Istanbul, Turkey: IEEE.
- Ngai, E. C., Liu, J., & Lyu, M. R. (2006). On the intruder detection for sinkhole attack in wireless sensor networks. In *IEEE International Conference on Communications* (pp. 3383-3389). Istanbul, Turkey: IEEE.
- Pirzada, A. A., & McDonald, C. (2005). Circumventing sinkholes and wormholes in wireless sensor networks. *IWWAN'05: Proceedings of International Workshop on Wireless Ad-hoc Networks*, 71.
- Polastre, J., Szewczyk, R., Mainwaring, A., Culler, D., & Anderson, J. (2004). *Analysis of wireless sensor networks for habitat monitoring*. In *Wireless sensor networks*. Springer.
- Qi, J., Hong, T., Xiaohui, K., & Qiang, L. (2012). Detection and defence of Sinkhole attack in Wireless Sensor Network. In *14th International Conference on Communication Technology* (pp. 809-813). Chengdu, China: IEEE.
- Rassam, M. A., Zainal, A., Maarof, M. A., & Al-Shaboti, M. (2012). A sinkhole attack detection scheme in minroute wireless sensor networks. In *International Symposium on Telecommunication Technologies* (pp. 71-75). Kuala Lumpur, Malaysia: IEEE. 10.1109/ISTT.2012.6481568
- Rehman, A., Rehman, S. U., & Raheem, H. (2019). Sinkhole Attacks in Wireless Sensor Networks: A Survey. *Wireless Personal Communications*, 106(4), 2291–2313. doi:10.1007/11277-018-6040-7
- S, A. I., & S, B. G. (2017). Cross Layer Approach For Detection and Prevention Of Sinkhole Attack Using A Mobile Agent. *2nd International Conference on Communication and Electronics Systems (ICES 2017)*, 359-365.
- Salehi, S. A., Razzaque, M. A., Naraei, P., & Farrokhtala, A. (2013). Detection of sinkhole attack in wireless sensor networks. In *IEEE International Conference on Space Science and Communication (IconSpace)* (pp. 361-365). Melaka, Malaysia: IEEE. 10.1109/IconSpace.2013.6599496
- Shafiei, H., Khonsari, A., Derakhshi, H., & Mousavi, P. (2014). Detection and mitigation of sinkhole attacks in wireless sensor networks. *Journal of Computer and System Sciences*, 80(3), 644–653. doi:10.1016/j.jcss.2013.06.016
- Sharmila, S., & Umamaheswari, G. (2011). Detection of sinkhole attack in wireless sensor networks using message digest algorithms. In *International Conference on Process Automation, Control and Computing* (pp. 1-6). Coimbatore, India: IEEE. 10.1109/PACC.2011.5978973
- Sheela, D., Kumar, C. N., & Mahadevan, G. (2011). A non cryptographic method of sink hole attack detection in wireless sensor networks. In *International Conference on Recent Trends in Information Technology (ICRTIT)* (pp. 527-532). Chennai, Tamil Nadu, India: IEEE. 10.1109/ICRTIT.2011.5972397
- Stafrace, S. K., & Antonopoulos, N. (2010). Military tactics in agent-based sinkhole attack detection for wireless ad hoc networks. *Computer Communications*, 33(5), 619–638. doi:10.1016/j.comcom.2009.11.006

Suryadevara, N. K., & Mukhopadhyay, S. (2012). Wireless sensor network based home monitoring system for wellness determination of elderly. *IEEE Sensors Journal*, 12(6), 1965–1972. doi:10.1109/JSEN.2011.2182341

Tumrongwittayapak, C., & Varakulsiripunth, R. (2009). *Detecting Sinkhole attacks in wireless sensor networks*. In ICCAS-SICE. IEEE.

Xiangning, F. &. (2007). Improvement on LEACH protocol of wireless sensor network. In *International Conference on Sensor Technologies and Applications* (pp. 260-264). Valencia, Spain: IEEE. 10.1109/SENSORCOMM.2007.4394931

Xu, N. (2002). A survey of sensor network applications. *IEEE Communications Magazine*, 40(8), 102–114. doi:10.1109/MCOM.2002.1024422

Zhang, R., & Xiao, X. (2019). Intrusion Detection in Wireless Sensor Networks with an Improved NSA Based on Space Division. *Journal of Sensors*, 2019, 1–20. doi:10.1155/2019/5451263

KEY TERMS AND DEFINITIONS

Energy Consumption: The amount of energy consumed using different methodologies in the proposed work

Hash Code Generation: A hashing algorithm is used to generate the hash code (an integer value) accepting the identification of the nodes as the input.

Public Key: Two keys are generated in public key cryptography. The public key is broadcasted by the source node to the nodes of the network. The nodes use the same to encrypt the messages before transmission. The source decrypts the received message using the private key.

Security Methodology: The measures used to bring in security to the nodes and the transmitted data in the network.

Signal Strength: The nodes based on their capacity transmit the packets with certain signal strength.

Sinkhole Attack: This is a kind of attack where the malicious nodes broadcast bogus messages to the nodes within the vicinity to bring them into confidence. The legitimate nodes transmit their packets to the malicious nodes which are later used to launch different kinds of attacks.

Wireless Sensor Network: Wireless sensor network (WSN) refers to a group of spatially dispersed and dedicated sensors deployed to monitor/track any object of interest.

TDSJ-IoT: Trivial Data Transmission to Sustain Energy From Reactive Jamming Attack in IoT

Ambika N.

 <https://orcid.org/0000-0003-4452-5514>

Department of Computer Application, SSMRV College, Bangalore, India

INTRODUCTION

Internet-of-things (Stankovic, 2014) is a combination of sensor technology with Radio Frequency Identification Device (RFID) (Jia, 2012) (Fan, Gong, Du, Li, & Yang, 2015), Geological Information System (GIS) (Liu, 2014) (Lopez, Lopez, Prieto, & Quinde, 2017) and Global positioning system (GPS) (Kumar, 2017) (Mala, Thushara, & Subbiah, 2017) (Jisha, Jyothindranath, & Kumary, 2017). The devices communicate with each other using integrated technology. The integration has aided to improve the quality of one's life minimizing human efforts. The technology consisting of a variety of devices aims to think, hear which aid in their doings and share information with each other. Four components are used to do the same. Sensing, heterogeneous access, information processing, applications, and services come together to bring the technology into play. The devices using the underlying technology turn the doings to become smarter. The applications include transportation (Melis, 2016) (Masek, et al., 2016), agriculture (Mat, 2016) (Popović, 2017), industry automation (Shrouf, 2014), emergency response (Yang, 2013) (Rathore, 2016) and healthcare (Muhammad, 2017) (Mano, 2016). Security (Andrea, 2015) and privacy components can be added to make the technology better.

Security in these devices is one of the prior issues to be considered as these devices will not be under continuous supervision. Hence the devices are liable to different kinds of attacks. Jamming attack (Alaba, 2017) (Kasinathan, 2013) is one such attack where the packets dispatched to the destination is hindered to reach the same. This leads to loss of packets and rarely the packets may reach the destination improperly. Many kinds of jammers are available. This includes basic jammers – reactive jammers, constant jammers, deceptive jammers, and random jammers. The second category of jammers namely intelligent jammers is of two kinds - statistical jammers and protocol-aware jammers. Each jammer has different characteristics and is active on different layers of the network. The descriptions of them are narrated in section 2.

The proposed work minimizes reactive jamming attack (Grover, 2014) (Babar, 2013) in the network. The work adopts some precautions to be taken to identify the adversaries before in hand. The methodology used will be able to detect the intruder earlier in the stage. Three kinds of precautions are taken in the proposed work. First, the devices are to sense the channel before commencing communication with any other device. They will be in continuous vigilance monitoring their surroundings. Second, to find the appropriate route, the device broadcasts sample bits addressing the receiving device. The reactive adversary sensing the packets are liable to transmit its packets. Hence the packets of the legitimate source will not reach the desired destination or will take a lot of time to reach its destination. The receiver affixes the time and path from which the packet it has received. Using this information the

DOI: 10.4018/978-1-7998-3479-3.ch038

further communication path is set. As the device is continuously sensing the channel it will be able to figure out jammers to some extent. Third, every device dispatching the packets will be equipped with device-id. Using the device-id, a hash code is derived which uniquely identifies the device. The hash code is prefixed to the transmitted packets. The receivers analyze the prefixed code along with signal strength and continuous with the communication if found legitimate. Using these steps reliability to the communication is increased by 4%, the effectiveness is increased by 7.6% and energy is minimized by 6.3% compared with the previous work.

The work is divided into eight sections. The introduction to the work is provided in section 1. Different kinds of jamming attacks are detailed in section 2 and Literature survey is explained in section 3. Section 4 contains the detailing of the proposed work. The details of the simulation are done in NS2 is provided in section 5 and the analysis of the same is provided in section 6. Section 7 details the future work to be taken. The work is concluded in section 8.

JAMMING ATTACKS

Different kinds of jamming attacks are available. Two kinds of jammers are available. They are basic and intelligent jammers (Hamza, 2016) (Noubir, 97-108). The basic jammers are into dispatching the packets are different intervals of time continuously or discreetly. Basic jammers are of four kinds- namely constant jammers, random jammers, deceptive jammers, and reactive jammers. The intelligent jammers are designed to understand the underlying technology and implement jamming to hinder the transmission. Intelligent jammers are of two kinds namely statistical jammer and protocol-aware jammer. Their characteristics are detailed in Table 1.

Table 1. Characteristics of different kinds of Jammers

Type of jammers		Characteristics
Basic jammers	Constant jammers (Han, 2018)	· This kind of jammer is effective on physical layer. · The jammer transmits continuous random bits without following any MAC label.
	Deceptive jammers (Lee, 2014)	· This kind of jammer is effective in MAC layer · The adversary continuously inject regular packets without any interval between packet transmission
	Random jammers (Wei, 2014) (Sanguanpong, 2018)	· The jammers toggle between sleep and jamming state. · This kind of jammer is effective in MAC layer
	Reactive jammers (Fadele, 2018) (Sciancalepore, 2018)	· This kind of jammer is effective in MAC layer · They emit radio signal on detecting any activity in the channel · These jammers keep the channel busy
Intelligent jammers	Statistical jammer (Ren, 2013)	· This kind of jammers are effective in physical layer · Observes the packet inter-arrival distribution and regulates itself to the traffic disrupting communication of the network
	Protocol-aware jammer (Sufyan, 2013) (Toledo, 2008)	· These kinds of jammers are effective in MAC layer. · It Keeps itself updated with the operating rules and deprives the legitimate nodes using its database.

BACKGROUND

Internet-of-things is used to link two devices and enables communication among them. As these devices are not under continuous supervision they are liable to different kinds of attacks. The proposed work aids in minimizing jamming attack in the network. The reactive jamming attack is considered in the work. Radio interference attacks are responsible to cause serious injury in the network. The following section provides a detailed study of the jamming attacks proposed by various authors.

In (Xu, 2005) jamming attacks are analyzed in wireless networks. The work is evaluated in two stages. In the first stage, the problems that crop up when radio interference attack is encountered is scrutinized. In the second stage, critical issues of diagnosing the presence of the attack are dissected. Four kinds of jamming attacks are introduced into the network and the effect on the wireless network, their ability to receive and send messages is evaluated. The work is analyzed considering the signal strength and carrier sensing time. The proposed work used Mica2 as the platform to conduct the respective experiments. The MAC protocol waits for the idle slot for a random amount of time to transmit the packets. The four kinds of jammer – constant jammer, reactive jammer, random jammer, and deceptive jammer were evaluated. Berkeley notes that employ ChipCon CC1000 RF transceiver and TinyOS operating system were used to implement and evaluate the working of the study under the four kinds of jamming attacks. The channels were disabled to sense the packets and the backoff operations to bypass MAC protocol to introduce the attacks in the network. The study will not be able to analyze the difference between the congested packet and jamming attack.

Reactive jamming is evaluated in (Wilhelm, 2011). An IEEE 802.15.4 network is used to examine the proposed work. The causes of loss at the physical layer are identified and a solution is provided to tackle reactive jamming attack. To detect the attack detection of jamming, scheduling and initialization of jamming and jamming burst to destroy the packet is calculated against the minimal time required for jamming. The proposed work concluded that minimal jamming duration is $26\mu\text{s}$, jamming initialization time was set to $15\mu\text{s}$ and the detector added a delay of less than $4\mu\text{s}$. The accounted data is used to check the same against USRP2-based reactive jamming prototype.

In (Raya, 2007) jamming attack was experimented on vehicular networks. Analyzing the threat analysis appropriate security architecture was devised. Digital signatures were used to protect the message legacy from the adversaries. The work adopts to use set of public/private keys pairs to digitally sign the messages and authenticate itself to the receivers. To tackle revocation of certificates a set of revocation protocols Revocation Protocol of the Tamper-proof Device (RTPD), Revocation protocol using Compressed Certificate Revocation list (RCCRL) and Distributed Revocation Protocol (DRP) are used. Using the protocol the malicious behavior can be detected.

A game-theoretic analysis is used to evade the attack of an aerial jammer on a communication channel (Bhattacharya, 2010). The kinematic model of the nodes is considered. To obtain optimal strategies, saddle point strategies are used. This is a zero-sum game which has a set of strategies for the players are used to arrive at the equations governing the saddle point strategies. In (Brown, James, & Sethi, 2006) the authors have considered an attacker disrupting an encrypted victim wireless adhoc network using jamming practice. Jamming at the Transport/network layer is considered in the work. The analysis revealed that TCP-SYN and TCP-SYN-ACK packets prevent TCP connection. ARP-REQUEST and ARQ-RESPONSE will prevent IP from associating IP and MAC address. Sensing is considered to be both online and offline. Online sensing packets are identified as they are received by the destination. Offline sensing is used to categorize the packets received in the history doings. The spread spectrum

proposed in the work provides a jamming immunity proportional to the spreading factor. The authors used Linux laptops running AODV-UU protocol while conducting testbed.

The authors in (Li M. K., 2007) have provided a solution for single-channel wireless sensor networks. The sophisticated jammer in the work controls the probability of the jamming and transmission range aims to corrupt the communication links in the network. The monitoring node is responsible to monitor the doings of the other nodes in the network. The supervising node conveys the alert message to the respective. The monitoring node is responsible to estimate the percentage of incurred collisions and draw out the outcome of optimal detection test. The network computes channel access probability to aid in minimizing jamming detection and reporting time. The study focuses to understand the problems regarding the structure, identifying the tradeoffs and capturing the impact of different parameters on its performance. The study is made to bring in a strategy which aims to handle the situation in case of knowledge-based/without a knowledge-based scenario. The network and monitor node without the knowledge uses the formulation and resolving optimization problems.

Path-based Dos attack is dealt in (Deng, 2005). The wireless sensor network will undergo serious damage. The intruder either injects replays the packets into the network or spurious packets. The adversary aims to block the path by introducing jamming packets and initiating the base station to re-boot the OHC (a sequence of numbers) for all the nodes on the path. Two methodologies are adapted aiding to minimize jamming and forgo the above scenario. OHC proactive bootstrapping employs to bootstrap the nodes near the jamming path and refreshing the OHC sequence. Another approach is known as lazy OHC bootstrapping where the addition of a new node to the network does not commence bootstrapping but still maintains resilience to stop Dos attack.

In (Hamieh, 2009) the authors have designed a solution for jamming attacks in mobile Adhoc networks. The authors have used a scenario where the jammer gets itself into its doings when valid radioactivity is signaled. The adversary moves to sleep state when it is not jamming the signals. To find the solution to this kind of attack, dependence among the periods of error and correct reception times are analyzed. The correlation coefficient is used to compute the same. The coefficient is the statistical measure of the relation between two random variables. To evaluate the jammer, error probability and correlation coefficient is measured.

In (Sampath, Dai, Zheng, & Zhao, 2007) 802.11 networks are considered. A detailed study on Single-channel jamming and multi-channel jamming attacks is done in this work. In the case of single-channel jamming attack, the adversary transmits high power signals on the channel. The impact of jamming depends on the size of the packets, length of inter-jamming interval and size of jamming packets. The adversary is observed to transmit 50-byte size packets to disrupt the data link. According to the experiment conducted in the work 100% —40% of packets suffer from degradation. Considering the second scenario, multiple radio devices are used to cause multi-channel jamming attacks. The impact can cause 5-100ms switching delay.

The solution to Spoofing-jamming attack in wireless smart grid network is detailed in (Gai, 2017). The work considers both jamming and spoofing to intrude in the communication. The frequencies used to transmit jamming packets vary with time. Spectrum adversarial occupations are used to maximize the attack effect. The study uses dynamic programming to solve power distribution optimal problem. The study is evaluated using the simulator MAS-SIM.

PROPOSED STUDY

Assumptions Made in the Study

- The base station is considered to be the most reliable resource.
- The network is vulnerable to jamming attack. The reactive jammer introduces the packets which it receives the signal from the channel when the channel transmits any of the packets.

Working of the Proposed Work

In (Xu, 2005) the author has suggested examining the signal strength and carrier sensing time. Analyzing the signal strength aids the receiver to accept/reject a packet from receiving. The adversary will usually aim to transmit the packets with high signal strength and hence scrutinizing the strength refrain from accepting the packets from the adversaries. Carrier sensing time aids in providing the picture of the occupancy of the channel. If the channel is constantly occupied the source will use another path to communicate its packets. The work proves to tackle jamming attack at its best in case of constant jamming. But the adversary introducing reactive jamming attack transmits its packets only when there is a packet movement in the channel. Hence using the methodology some initial packets can get lost and the source may have to retransmit the same.

The proposed work is suggested to improve the work suggested in (Xu, 2005). The work is put forward to tackle the reactive jamming attack. In this kind of attack, the adversary transmits the packets when it senses any transmission in the channel. The proposed work operates using the following steps-

- The devices are habituated to sense the channel before commencing communication with any other device. They are in continuous vigilance monitoring their surroundings.
- To find the appropriate route, the device broadcasts sample bits addressing the receiving device. The reactive adversary sensing the packets are liable to transmit its packets. Hence the packets of the legitimate source will not reach the desired destination or will take a lot of time to reach its destination. The receiver affixes the time and path from which the packet it has received. Using this information the further communication path is set. As the device is continuously sensing the channel it will be able to figure out jammers to some extent.
- Every device dispatching the packets will be equipped with device-id. Using the device-id, a hash code (Riedel, 2010) is derived which uniquely identifies the device. The hash code is prefixed to the transmitted packets. The receivers analyze the prefixed code along with signal strength and continuous with the communication if found legitimate.

SIMULATION OF THE WORK

The proposed work considers three steps to tackle reactive jamming attack. The devices are under continuous vigilance. They try to be completely aware of their surroundings by listening to the channel before any transmission. This procedure keeps them aware of the jamming attack in its vicinity. Secondly, they transmit sample data to probe the adversary into its act. Hence the methodology aids in locating the jammer position and keeps its packets away from them. To enhance reliability to the transmitted packets, the devices prefix hash code generated by the device-id.

The work is simulated using NS2. MICA2 motes (Narayanan, Sarath, & Vineeth, 2016) are considered in the simulation work. 4 motes are installed in the environment of area 200m * 200m. The sample bits are of 4 bits in length. The hash code is 8 bits in length and the transmitted data is of 64 bits in length. The work is simulated for the 60s. Table 2 depicts the details of the simulation.

Table 2. Depiction of simulation parameters used in the study

Parameters used	Description
Simulation area	200m * 200m
Number of motes installed	4
Number of jammers	1
Length of sample bits	4
Length of hash code	8 bits
Length of transmitted bits	64 bits
Simulation time	60 s

ANALYSIS OF THE PROPOSED WORK

Energy Consumption

Energy (Al-Ali, Zualkernan, Rashid, Gupta, & Alikarar, 2017) (Chiu, Sun, Thompson, Nakayama, & Zhang, 2017) is one of the most important resources in these devices. The devices lose a certain amount of energy while receiving and accepting packets. Hence to curtail the energy consumption by avoiding retransmission and accepting legitimate packets the proposed work is suggested.

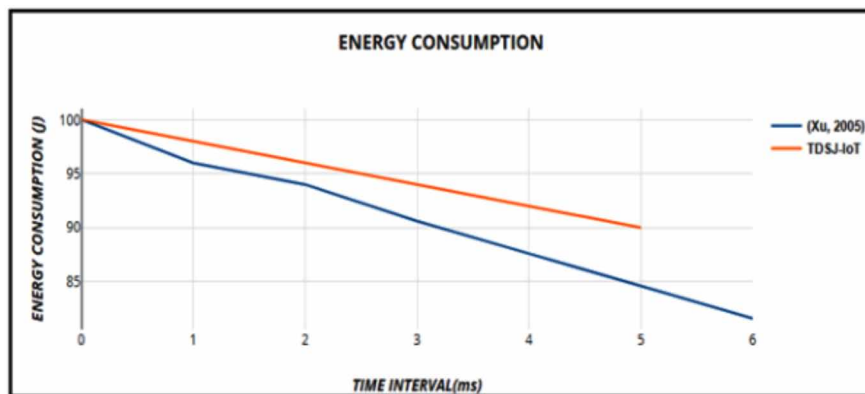
The proposed work aids in minimizing the energy consumption of the receiving device. Three steps are considered in the proposed work to minimize energy getting drained in the receiving node. The energy is curtailed as the devices become aware of their environment and can transmit their packets safely to their respective destination. The methodology restricts retransmission of the packets.

- The devices are into the practice of examining and sensing the channel before commencing any kind of transmission.
- An appropriate path is chosen to communicate with the other device by using the broadcasting mechanism.
- The packets are prefixed with the hash code of the respective device providing an identity of the packet origin.

The work suggested by (Xu, 2005) uses signal strength to analyze the packets. The adversaries are assumed to transmit the packets with high signal strength and hence only the packets with appropriate signal strength are accepted. As the network is dealing with reactive attacks, the adversary transmits its packets only when it senses packets in the channel. The suggestion provided by (Xu, 2005) would not be able to trace the adversary on time and hence will suffer from loss of some of its packets in the attack. This triggers the source to retransmit the packets to the destination. The proposed work will be able to detect the attack while it broadcast sample bits into its environment and sensing the channel. Hence

considerable numbers of packets are saved from the loss against jamming attack. This, in turn, consumes some amount of energy in these devices. Comparing the proposed work with (Xu, 2005) nearly 6.3% of energy is saved using three energy conservation steps. The same is depicted in Figure 1.

Figure 1. Depiction of energy consumption in TDSJ-IoT and (Xu, 2005)

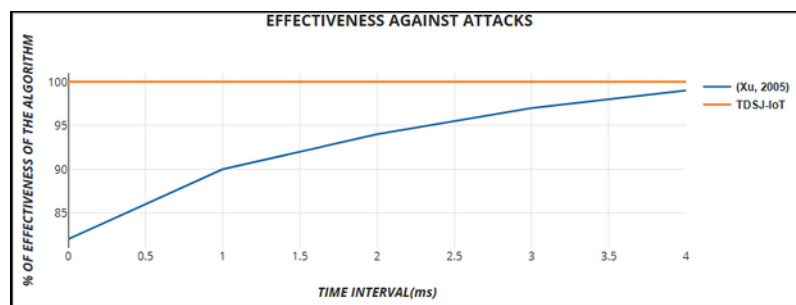


Effectiveness of the Steps Considered in the Work

Carrier sensing time and signal strength are analyzed in (Xu, 2005) to transmit the packets into the channel during that duration of time. The work proves to work better with the constant jamming attack but will shortfall with the reactive jamming attack.

The proposed work analyses the channel in multiple ways which aid the devices to conserve some amount of energy and also tackles the attack on time. The work aids in monitoring the environment continuously to bring in better awareness of the vicinity to the devices. Using sample bits in transmission triggers the adversary and keeps the legitimate devices aware of their location. Prefixing the hash code to the transmitted packets enables to improve reliability to the transmission. Comparing the work suggested in (Xu, 2005) and the proposed work proves to be effective by 7.6%. Figure 2, provides the depiction of the comparison of the effectiveness of the proposed work and the work suggested in (Xu, 2005).

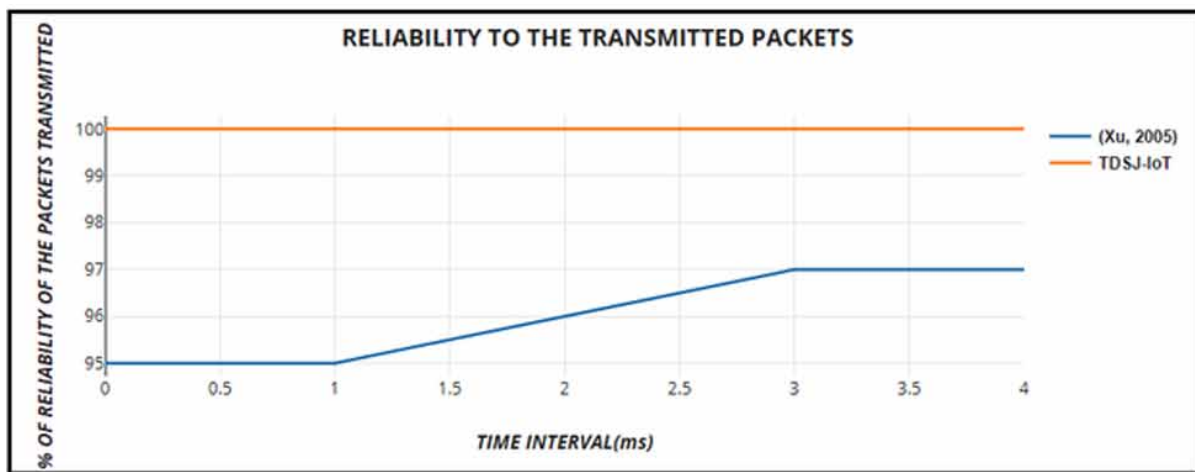
Figure 2. Graphical Representation of effectiveness of the algorithm against attacks



Reliability to the Transmitted Packets

In (Xu, 2005) the authors have suggest minimizing the jamming attack by using carrier sense time and signal strength. The proposed work adds another vital ingredient to the transmission. The packets received by the destination will provide more reliability (Li L. J., 2012) (LI, 2012) (Ahmad, 2014). The hash code is generated using the device-id and this code becomes the identification of the device and its transmitted packets. Comparing the work with (Xu, 2005) the proposed work provides 4% more reliability to the transmitted packets. Figure 3 provides the graphical representation of the comparison of the proposed work with (Xu, 2005) concerning the reliability of the transmitted packets by the source to the respective destination.

Figure 3. Depiction of reliability of packet transmission in reactive jamming prone network



FUTURE WORK

The proposed work aids in minimizing the reactive jamming attack in the environment. The challenge is how it secures its packets from the jammer. It follows three steps to do the same. It will be under continuous vigilance monitoring its vicinity. This methodology aids in making the devices aware of their environment. The devices trigger the adversary by broadcasting sample bits in the network. This enables the devices to locate the jammers and keep their packets away from them. To increase reliability to the transmitted packets it prefixes hash code to the packets before transmission. To make the methodology even better the following points are to be considered-

Energy Consumption

As the devices are into continuous awareness, the energy consumption is relatively high comparative to devices which move into a sleep state. Hence better methodologies have to be devised to adopt some duration of sleep to these devices. The devices broadcast sample bits every time it has to initiate communication. This methodology also consumes a lot of energy and has to be curtailed.

Change of Hash Code

The hash code derived from the device-id and prefixed to the transmitted packets remains the same throughout. Hence to keep the transmission safer better methodology has to be adopted to change the hash code with different sessions.

CONCLUSION

Internet-of-things is a new technology which enables many devices to communicate with each other. The devices in communication may belong to the same category sharing the same platform or may belong to different categories. The technology aids in bridging the gap between the categories and building a common platform where the devices will be able to communicate with each other securely. As the devices are not under supervision, they are liable to different kinds of attacks. This paper tackles reactive attacks in the network. Energy and timeliness are some of the prior issues to be addressed in this network.

The proposed work uses three effective steps to minimize the packets getting lost or destroyed in the jamming attack. The algorithm is into sensing the channel which keeps the devices updated with the surrounding environment it is residing at that instant of time. The second suggestion provided by the proposed work is to broadcast bits of sample data. Utilizing this methodology the adversaries are triggered to transmit their jamming packets to hinder transmission. Hence the adopted methodology aims to identify the location of the jammer. The proposed work using the methodology will use a reliable path to transmit its packets to the destination. To make the transmission more reliable, a hash code generated using the unique identification of the device. This, in turn, enhances reliability to the packets delivered. The proposed work proves to be more effective by 7.6% than the previous work and also conserves 6.3% of energy. Reliability of the transmission of packets is increased by 4%.

REFERENCES

- Ahmad, M. (2014). Reliability Models for the Internet of Things: A Paradigm Shift. In *IEEE International Symposium on Software Reliability Engineering Workshops*. Naples, Italy: IEEE. 10.1109/ISSREW.2014.107
- Al-Ali, A., Zuolkernan, I. A., Rashid, M., Gupta, R., & Alikarar, M. (2017). A smart home energy management system using IoT and big data analytics approach. *IEEE Transactions on Consumer Electronics*, 63(4), 426–434. doi:10.1109/TCE.2017.015014
- Alaba, F. A., Othman, M., Hashem, I. A. T., & Alotaibi, F. (2017). Internet of Things security: A survey. *Journal of Network and Computer Applications*, 88, 10–28. doi:10.1016/j.jnca.2017.04.002
- Andrea, I. C. (2015). Internet of Things: Security vulnerabilities and challenges. In *IEEE Symposium on Computers and Communication (ISCC)*, (pp. 180-187). IEEE. 10.1109/ISCC.2015.7405513
- Babar, S. D. (2013). Jamming attack: Behavioral modelling and analysis. In *International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE)* (pp. 1-5). IEEE.

- Bhattacharya, S. B. (2010). Game-theoretic analysis of an aerial jamming attack on a UAV communication network. *American Control Conference* (pp. 818-823). Baltimore, MD: IEEE. 10.1109/ACC.2010.5530755
- Brown, T. X., James, J. E., & Sethi, A. (2006). Jamming and sensing of encrypted wireless ad hoc networks. In *7th ACM international symposium on Mobile ad hoc networking and computing* (pp. 120-130). Florence, Italy: ACM.
- Chiu, W.-Y., Sun, H., Thompson, J., Nakayama, K., & Zhang, S. (2017). IoT and Information Processing in Smart Energy Applications. *IEEE Communications Magazine*, 55(10), 44. doi:10.1109/MCOM.2017.8067682
- Deng, J. H. (2005). Defending against path-based DoS attacks in wireless sensor networks. In *3rd ACM workshop on Security of ad hoc and sensor networks* (pp. 89-96). Alexandria, VA: ACM. 10.1145/1102219.1102235
- Fadele, A. A. (2018). A novel countermeasure technique for reactive jamming attack in internet of things. *Multimedia Tools and Applications*, 1–22.
- Fan, K., Gong, Y., Du, Z., Li, H., & Yang, Y. (2015). *RFID Secure Application Revocation for IoT in 5G*. In *IEEE Trustcom/BigDataSE/ISPA*. IEEE.
- Gai, K. Q., Qiu, M., Ming, Z., Zhao, H., & Qiu, L. (2017). Spoofing-jamming attack strategy using optimal power distributions in wireless smart grid networks. *IEEE Transactions on Smart Grid*, 8(5), 2431–2439. doi:10.1109/TSG.2017.2664043
- Grover, K. L., Lim, A., & Yang, Q. (2014). Jamming and Anti-jamming Techniques in Wireless Networks: A Survey. *International Journal of Ad Hoc and Ubiquitous Computing*, 17(4), 197–215. doi:10.1504/IJAHUC.2014.066419
- Hamieh, A. B.-O. (2009). Detection of jamming attacks in wireless ad hoc networks using error distribution. In *IEEE International Conference on Communications* (pp. 1-6). Dresden, Germany: IEEE. 10.1109/ICC.2009.5198912
- Hamza, T. K. (2016). A survey on intelligent MAC layer jamming attacks and countermeasures in WSNs. In *84th Vehicular Technology Conference* (pp. 1-5). IEEE. 10.1109/VTCFall.2016.7880885
- Han, G. L., Liu, L., Zhang, W., & Chan, S. (2018). A hierarchical jammed-area mapping service for ubiquitous communication in smart communities. *IEEE Communications Magazine*, 56(1), 92–98. doi:10.1109/MCOM.2018.1700399
- Jia, X. F. (2012). RFID technology and its applications in Internet of Things (IoT). In *2nd International Conference on Consumer Electronics, Communications and Networks*, (pp. 1282-1285). Yichang, China: Academic Press. 10.1109/CECNet.2012.6201508
- Jisha, R., Jyothindranath, A., & Kumary, L. S. (2017). Iot based school bus tracking and arrival time prediction. In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Udupi, India: IEEE. 10.1109/ICACCI.2017.8125890
- Kasinathan, P. P. (2013). Denial-of-Service detection in 6LoWPAN based Internet of Things. In *IEEE 9th international conference on wireless and mobile computing, networking and communications (WiMob)* (pp. 600-607). IEEE.

- Kumar, S. H. (2017). IoT approach to save life using GPS for the traveller during accident. In *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, (pp. 2424-2428). Chennai, India: IEEE. 10.1109/ICPCSI.2017.8392152
- Lee, T. H. (2014). A lightweight intrusion detection scheme based on energy consumption analysis in 6LoWPAN. *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, 1205-1213.
- Li, W. M. (2012). IOT System Reliability Testing and Evaluate Technology. *Software*, 1.
- Li, L. J. (2012). Modeling and analyzing the reliability and cost of service composition in the IoT: A probabilistic approach. In *IEEE 19th International Conference on In Web Services* (pp. 584-591). IEEE.
- Li, M. K. (2007). Optimal jamming attacks and network defense policies in wireless sensor networks. In *26th IEEE International Conference on Computer Communications* (pp. 1307-1315). Barcelona, Spain: IEEE. 10.1109/INFCOM.2007.155
- Liu, S. J., & Zhu, G. (2014). The application of GIS and IOT technology on building fire evacuation. *Procedia Engineering*, 71, 577–582. doi:10.1016/j.proeng.2014.04.082
- Lopez, D. A., Lopez, J. R., Prieto, M. A., & Quinde, L. D. (2017). Towards a Method for the Integration of IoT and GIS Applications Deployed on Cloud Platforms. In *International Conference on Information Systems and Computer Science (INCISCOS)*. Quito, Ecuador: IEEE.
- Mala, N. S., Thushara, S. S., & Subbiah, S. (2017). Navigation gadget for visually impaired based on IoT. In *2nd International Conference on Computing and Communications Technologies (ICCCT)*. Chennai, India: IEEE. 10.1109/ICCCT2.2017.7972298
- Mano, L. Y., Façal, B. S., Nakamura, L. H. V., Gomes, P. H., Libralon, G. L., Meneguete, R. I., Filho, G. P. R., Giancristofaro, G. T., Pessin, G., Krishnamachari, B., & Ueyama, J. (2016). Exploiting IoT technologies for enhancing Health Smart Homes through patient identification and emotion recognition. *Computer Communications*, 89, 178–190. doi:10.1016/j.comcom.2016.03.010
- Masek, P., Masek, J., Frantik, P., Fujdiak, R., Ometov, A., Hosek, J., Andreev, S., Mlynek, P., & Misurec, J. (2016). A Harmonized Perspective on Transportation Management in Smart Cities: The Novel IoT-Driven Environment for Road Traffic Modeling. *Sensors (Basel)*, 16(11), 1872. doi:10.3390/16111872 PMID:27834796
- Mat, I. K. (2016). IoT in precision agriculture applications using wireless moisture sensor network. In *IEEE Conference on Open Systems (ICOS)*, (pp. 24-29). Langkawi, Malaysia: IEEE. 10.1109/ICOS.2016.7881983
- Melis, A. P. (2016). Public transportation, IoT, trust and urban habits. *International Conference on Internet Science*, 318-325. 10.1007/978-3-319-45982-0_27
- Muhammad, G. R., Rahman, S. K. M. M., Alelaiwi, A., & Alamri, A. (2017). Smart health solution integrating IoT and cloud: A case study of voice pathology monitoring. *IEEE Communications Magazine*, 55(1), 69–73. doi:10.1109/MCOM.2017.1600425CM
- Narayanan, R. P., Sarath, T. V., & Vineeth, V. V. (2016). Survey on Motes Used in Wireless Sensor Networks: Performance & Parametric Analysis. *Wireless Sensor Network*, 51-60.

- Noubir, G. R. (97-108). On the robustness of IEEE 802.11 rate adaptation algorithms against smart jamming. In *Fourth ACM conference on Wireless network security* (p. 2011). ACM. 10.1145/1998412.1998430
- Popović, T. L., Latinović, N., Pešić, A., Zečević, Ž., Krstajić, B., & Djukanović, S. (2017). Architecting an IoT-enabled platform for precision agriculture and ecological monitoring: A case study. *Computers and Electronics in Agriculture*, 140, 255–265. doi:10.1016/j.compag.2017.06.008
- Rathore, M. M., Ahmad, A., Paul, A., Wan, J., & Zhang, D. (2016). Real-time medical emergency response system: Exploiting IoT and big data for public health. *Journal of Medical Systems*, 40(12), 283. doi:10.1007/10916-016-0647-6 PMID:27796839
- Raya, M. H., & Hubaux, J.-P. (2007). Securing vehicular ad hoc networks. *Journal of Computer Security*, 15(1), 39–68. doi:10.3233/JCS-2007-15103
- Ren, K. (2013). Opportunistic spectrum access: From stochastic channels to non-stochastic channels. *IEEE Wireless Communications*, 20(3), 128–135. doi:10.1109/MWC.2013.6549292
- Riedel, T. F. (2010). Using web service gateways and code generation for sustainable IoT system development. *Internet of Things*, 1-8.
- Sampath, A., Dai, H., Zheng, H., & Zhao, B. Y. (2007). Multi-channel Jamming Attacks using Cognitive Radios. In *16th International Conference on Computer Communications and Networks* (pp. 352-357). Honolulu, HI: IEEE. 10.1109/ICCCN.2007.4317844
- Sanguanpong, S. (2018). Secrecy Outage Performance Analysis for Energy Harvesting Sensor Networks With a Jammer Using Relay Selection Strategy. *Special Section on Security and Trusted Computing for Industrial Internet of Things*, 23406-23419.
- Sciancalepore, S. O., Oligeri, G., & Di Pietro, R. (2018). Strength of Crowd (SOC)—Defeating a Reactive Jammer in IoT with Decoy Messages. *Sensors (Basel)*, 18(10), 3492. doi:10.3390/18103492 PMID:30332848
- Shrouf, F. O. (2014). Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm. In *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, (pp. 697-701). Bandar Sunway, Malaysia: IEEE. 10.1109/IEEM.2014.7058728
- Stankovic, J. A. (2014). Research directions for the internet of things. *IEEE Internet of Things Journal*, 1(1), 3–9. doi:10.1109/JIOT.2014.2312291
- Sufyan, N. S., Saqib, N. A., & Zia, M. (2013). Detection of jamming attacks in 802.11 b wireless networks. *EURASIP Journal on Wireless Communications and Networking*, 2013(1), 208. doi:10.1186/1687-1499-2013-208
- Toledo, A. L., & Xiaodong Wang. (2008). Robust detection of MAC layer denial-of-service attacks in CSMA/CA wireless networks. *IEEE Transactions on Information Forensics and Security*, 3(3), 347–358. doi:10.1109/TIFS.2008.926098
- Wei, X. H. (2014). A jammer deployment method for multi-hop wireless network based on degree distribution. In *Ninth International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)* (pp. 261-266). IEEE. 10.1109/BWCCA.2014.73

Wilhelm, M. M. (2011). Short paper: reactive jamming in wireless networks: how realistic is the threat? In *Fourth ACM conference on Wireless network security* (pp. 47-52). Hamburg, Germany: ACM. 10.1145/1998412.1998422

Xu, W., W. T. (2005). The Feasibility of Launching and Detecting Jamming Attacks in Wireless Networks. In *6th ACM International Symposium on Mobile Ad Hoc Networking and Computing - MobiHoc*, (pp. 46-57). ACM. 10.1145/1062689.1062697

Xu, W. T. (2005). The feasibility of launching and detecting jamming attacks in wireless networks. In *6th ACM international symposium on Mobile ad hoc networking and computing* (pp. 46-57). ACM.

Yang, L. Y., Yang, S. H., & Plotnick, L. (2013). How the internet of things technology enhances emergency response operations. *Technological Forecasting and Social Change*, 80(9), 1854–1867. doi:10.1016/j.techfore.2012.07.011

KEY TERMS AND DEFINITIONS

Hash Code: Hash code is a numeric value generated that helps in identifying the object.

Internet of Things: The Internet of Things shortly known as IoT is a system of interrelated computing devices. The objects of interest are provisioned with unique identifications. This facility aids in transfer data over a network without human-to-computer interaction.

Reactive Jamming Attack: This kind of jammer is effective in MAC layer. The adversary emits radio signal on detecting any activity in the channel. These jammers keep the channel busy.

Section 5

Data Management, Analysis, and Statistics

Software Component Technology: Concepts, Design, and Management Method

Fadoua Rehioui

 <https://orcid.org/0000-0001-6814-6717>

University Soultan Moulay Slimane, Morocco

INTRODUCTION

Data and Knowledge are at the level of all organizations. Due to the amount of data generated, companies require different analyzes, supporting data-driven decision making (Nimmy et al., 2017). Well-organized management of data and knowledge are the key to the success science related data and then the almost of any enterprise.

The issue of data and knowledge management (Ermine and al., 1996) has been defined as the implementation of a management system that allows all components of the company (Brunet et al., 1994) to use and enrich knowledge and to consider this heritage of knowledge as the same subject of the information system of the company (Ermine, 2000). As a result, the technological revolution has brought about a change in the way information/data is collected (Kamal et al., 2017) and treated and so the implication of the *Components*.

Any conventional design of complex software systems always starts with the identification of the system in parts (subsystems, blocks, class, etc.) where the idea and the concept of the components come from, i.e. the decomposition of the system is to bring out the parts which are interested. It is no coincidence that the notion of component gives a great interest in the development of complex systems.

Certainly, the importance and usefulness of components in scientific research are not new topics, especially in the reuse of components. Formerly, this work focuses on the complete definition of component-based information systems approach design.

The idea of component-based development is introduced by defining the concept and providing its management. The approach primarily seeks to explain and analyze and manage the variety of data and information exchanged in a company. This method is useful for a global vision of the company's system, using phases and steps designed tailored to the needs. This represents a lot of "Components" not yet developed in implementation.

The difficulty is to design the components and the services most relevant to the goals. Design requires a method for modeling that involves creating a virtual representation in such a way as to bring out the goals interested in and build upon.

In fact, there is no standard for modeling a component-based system, so it is necessary to try to propose a component identification approach, and an approach to illustrate the management of a software system.

The objectives of this chapter the contribution of a new solution of the method based on software components, and a communication management approach between these identified software components mediators between the components and the information system. Hence, the idea that software systems can be developed by components intended to provide specific services such as combining, managing, connecting and communicating between system units.

DOI: 10.4018/978-1-7998-3479-3.ch039

This paper is organized in separate paragraphs and sections, distributed as follows:

- Section 2: Provides definitions incorporating general discussions on Data Management and Component Development, presents the state of the art (literature review).
- Section 3: formalizes the model of the proposed approaches.
- Section 4: presents the different elements of the technology served in this area.
- Conclusion: The paper ends with a conclusion.

BACKGROUND (SECTION 2)

Commonly, a system processes the data with a set of actions from a huge store of information and then invoking the input / output functions. A computer system has the function, in a very general way, of processing information. One of the first difficulties is to define the notion of information and in particular to delimit it in relation to related notions such as data and knowledge. Knowledge is used all the time and on many occasions, it refers to information. The difficulty of defining knowledge comes from its relationship to two concepts, data and information. Data are not organized in any way and provide no further information. As a result, data, information and knowledge are defined as follows:

The first concept to meet is the **data** which is a raw element not yet explained or guessed. The contextualization in an interpreted **data**, by creating one or more added values, generates **information** (Davenport & Prusak 2000).

Information systems are supported by processing systems that process only data and then the **data** is transformed into **information** when it is communicated to a human being capable of interpreting and deciding it (Bali et al. 2009).

The interpreted **information** is then transformed into contextualized **knowledge** and put in relation with existing **knowledge**.

Subsequently, **knowledge** comes and considers itself as **information** understood, that is to say, assimilated and used, which leads to an action (Davenport & Prusak 2000).

For **data** to become **information**, it is required to be contextualized, categorized. The **data** give the key to make the decisions and methods suitable for improvement (Kamal and al., 2016), supporting data-driven decision-making and management.

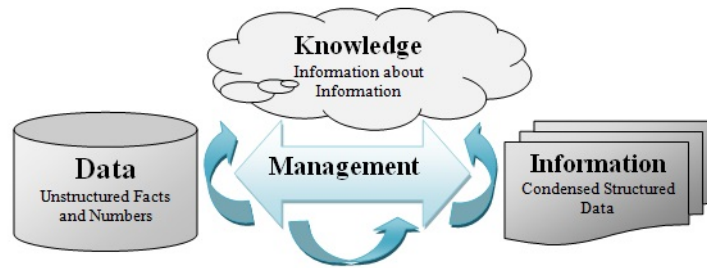
Design of meaning assigned to the data to make it an information and knowledge in a Micro-credit Agency does not extend credits for people over 60 years of age: the **data** 05-11-1956 generates the **information** that November 5, 1956 is a date of birth and 64 years is the **knowledge** that will alert and decide that his file will be rejected by the system.

In order for Knowledge Management (KM) to succeed, one needs a deep understanding of what constitutes knowledge.

More technologically, it is often treated as something can be codified and transmitted, similarly to information mainly involving information systems- knowledge. These two terms are often regarded as lower denominations of knowledge. However, even today, some KM systems are little more than information management systems using knowledge as a virtual synonym for information. Here, in summary the meeting of the three concepts:

- **Data** is the lowest point, an unstructured collection of facts and numbers;
- **Information** is the next level, and it is regarded as structured data;
- Finally **knowledge** is defined as “information about information”.

Figure 1. The relationship of the Management process: Data - Information- Knowledge



As a result of the amounts of data generated (Nimmy et al., 2017), companies require different analyzes, supporting data-driven decision-making, as enterprise resource planning (ERP) systems handle a huge amount of data, related to the actual execution of business processes and the work of (Pajić and al., 2016) is a detailed analysis of the concepts of approach centered on the artifact and describe its constructions by the meta-model and the capture of knowledge in data conversion. As a result, the technology revolution has brought a change in the way data are collected (Kamal et al., 2017).

In practice, information systems collect data by organizing them in easily accessible parts, but it is rare to directly define the status of implementation of dependencies between its different parts. The study of the treatment of multiple parts is generally simpler than the study of the overall treatment of a system, because each simpler component will be the subject of a separate study and simpler to study separately.

The principle of developing systems from the components is not new, it is even the object approach is based on the initial identification of the different elements of the system, then, modules and classes corresponding to each element. Irregularly, the term component is simply used as an equivalent for the object-oriented architecture.

The discussion in this paper is on some component-based approaches to object-oriented approaches, such as: UML Component (Cheesman, 2001) and Extraction of Component-Based Architecture from an Object-Oriented System (Sylvain et al., 2007), and others related to software-based engineering.

The approach, according to this work must start from the general, and go towards the particular, that is to say a guided by the functionalities of the system; the system is divided into components where each must perform a function (possibly several, if the component is decomposed into subcomponents with separate functions). These components make it possible to organize and cut the system into manageable and exchangeable pieces of software. So, what is a component and what is its task..

The idea of the component was first proposed in 1968 by (Douglas McIlroy, 1968). Before the software was built in one piece, and not assembled from components. With this lack of software component creation and modeling, software engineering has shifted the trend of application and software composition from separately developed software segments (Sosnin.P, 2016).

In computer science, a database can be component oriented, i.e. a database containing the set of data stored under a data management model, and programming via database management systems (Karaa et al., 2016).

The Component

The component is essentially characterized by two elements: its interfaces and its properties. An interface is a “use / provide” dependency relationship between the component and its environment. These

interfaces are characterized by the methods they define and by a direction of communication (provided, incoming / requested, outgoing) and can be defined as a specification of its access points via the port.

Between port and component, a connector connects these two or more connectable elements. It can be “Delegation” or “Assembly”. A delegation connector links a component to its internal subcomponents (Accord, 2002).

Typically, a component interface represents a set of functional and non-functional properties that are outside that component. The so-called functional properties describing the structure, the behavior and the functionalities of the component and the non-functional properties concerning the services used by the component that are not part of the application services (security, performance, etc.).

The Composite Component

The components are divided into two types, a primitive component and / or a composite component whose internal architecture is formed by several components, exposing its contents, named subcomponents, and are instantiated, interconnected and configured with the composite; represent a role played by a set of instances at runtime.

To avoid a common confusion of Component and Object, are they different things or not. Several temptations and directions have used the object to develop an optional component.

Literature Review

This part is dedicated to the introduction of component-based approaches related to object-oriented approaches.

• From Object Oriented to Component Oriented

The contribution of the components to the object is notably a greater modularity thanks to the encapsulation of the data and the explicit coupling of the dependencies required and provided. The classical approaches designate the notion object by the definition and the assembly of the element objects grouped the data and the treatments of these data.

Irregularly, the term component is simply used as an equivalent for the object-oriented architecture. Objects have a unique and inseparable class identity that encapsulates state, structure, and behavior. These classes are used to create a copy of a class named object (an instance of the class), and provide the implementation of the object behavior. As for the component, it breaks down into an outer part for the interfaces and another part /of the implementation. It is a storage unit to which is associated an implementation unit, allowing the description of the internal operation. In fact, a component can be implemented by a single class or by multiple classes, often hidden in the implementation and sometimes only available in binary form. The software components agree with the standards defined by a component model, they can come in different forms: class, specification, documentation, source code, class hierarchy, interface, etc.

a. UML Component (2001)

J.Cheesman and J.Daniels have proposed the UML Component method (Cheesman, 2001), for the design, development Objects, inspired by several methods, such as Catalysis (D’souza and al., 1998)

and RUP (Kruchten, 1998). UML Components, as the name implies, applies UML in the component architecture, to specify the components, their interactions, i.e. focuses on component specification by enriching UML notation with stereotypes to designate the specification of system models and divide these models into components.

Throughout, the emphasis is placed on the specification of the external characteristics of the components and their interdependencies, rather than on the internal activity. The study of the interactions between components is defined by the interfaces and therefore by calls of methods and services between components.

The proposed method is limited to the design level. Indeed, it does not explain how to interpret the specification of the component, in the implementation, or how to check the specifications at the compilation level.

b. Extraction d'Architecture à Base de Composants d'un Système Orienté Objet, 2007

Sylvain Chardigny, Abdelhak Seriai, Mourad Oussalah and Dalila Tamzalit (Sylvain and al., 2007) have developed an approach to extract a component-based architecture for the implementation of an existing object-oriented system. The primary idea of this approach is to propose a semi-automated process of identifying the components of a system based on their semantic and structural characteristics.

Extracting architecture is the reverse process of its design. Instead of using the architect's knowledge to create architecture and implement it, the extraction process uses its knowledge to obtain an abstraction of the system from its implementation.

c. Identifying a Component-Based Architecture in an Object-Oriented Application Using Dynamic Analysis (2011)

Systems with large numbers of elements, describes a complexity difficult to understand and maintain. The component paradigm provides a good representation of high-level complex systems, and can be used as "projection" of object-oriented systems.

Therefore, the goal of this work (Allier and al., 2011) is to propose a semi-automatic method to identify a component-based architecture in an object-oriented application.

The method should help in understanding the original application, as well as simplifying the projection of the application into a concrete component model.

The identification of a component-based architecture is done in three main steps:

- The collection of essential data for the identification process, corresponding to the dependencies between the classes and arrived at from a dynamic analysis of the application.
- Identification of the components, in three methods explored: the first one uses a lattice of Galois, the second two meta-heuristics and the last one a meta-heuristic multi-objective.
- Identification of the architecture based on application components, identifying the required / provided interfaces of each component.

The concept of projecting this component-based method in an object-oriented application; can be used as a template to perform a restructuring of an object-oriented application to a similar, component-based, operational application.

d. A formal approach to automate change management in component-based software development processes (2015)

This approach (Mokni, 2015) improves the management of software evolution in component-oriented development processes, adopting a model-driven engineering (IDM) approach, supported by *Dedal*, an architectural description language (ADL), three levels (specification, configuration and assembly) correspond to the three main stages of the development of an architecture (design, implementation, deployment) and keep the signature of the architectural decisions made during development.

This efficient management of the evolution makes it possible to determine the level of a change, to analyze its impact and to plan its realization in order to avoid the occurrence of inconsistencies in the definition of the architecture. Rigorous management of the evolution requires the formalization, on the one hand, of the intra-level relations linking the components within the models corresponding to the different levels of definition of architecture and, on the other hand, inter-level relations linking models describing the same architecture to different levels of abstraction. Relationships and properties are described in B, a formal modeling language based on set theory and first-order logic. They apply to architectures defined with a formal ADL written in B whose meta-model is aligned with that of *Dedal*. This integration makes it possible to propose a development environment combining the advantages of IDM approaches and formal design of architectures with *Dedal* tooling (graphical modeler); Architectural Verification and Evolution Management with B Tooling.

Knowledge Management Approaches

This part is devoted to some management knowledge approaches

a. Modeling behaviors in dynamic systems: Application to the organization and regulation of parking and travel in Urban Traffic Systems (2007)

This work (Boussier, 2007) proposes a methodological framework to model the behavioral changes of individuals as well as their impacts on the state of a complex system in a dynamic context.

Modeling the heterogeneous preferences of users, integrating knowledge, according to different behavioral logics, justified the multi agent approach, in order to have adaptive and evolutionary representations of behaviors.

Individual agents, decision-making and learning processes are modeled using a basket for tools from the optimization of industrial processes, Data Mining techniques, marketing. The personalized use of the simulator with simulator user interactivity motivated the adaptation of this methodological framework to the simulation of complex societal systems.

b. MODELS AND TOOLS FOR CAPITALIZATION OF KNOWLEDGE IN DESIGN: Contribution to management and knowledge engineering at Renault - DCT (2011)

The contribution of this work (Ludovic, 2011) is positioned within the scientific disciplines of the knowledge management and knowledge engineering.

From a theoretical point of view, the analysis of the state of the art, but also the context particular industry in which this work was carried out, led to propose a conceptual model of process-oriented

knowledge preservation. This target model mainly organizations, having chosen a global management approach by the process as recommended by ISO (International Organization of Standardization).

It is based on an original arrangement of scientific concepts strongly inspired by the method MASK (Methodology for Analysis and Structuring Knowledge) developed by Jean-Louis ERMINE (Aries, 2008) and having contributed to the preservation of knowledge . The interest of this model is positioned in a goal to manage, by through a formalization stage, the link that exists between, on the one hand, codified knowledge in documents or knowledge-based applications and on the other hand concrete operational activities. This management is targeted optimizing the quality, costs and delays of redesign processes.

FOCUS OF THE ARTICLE

Essentially, the heart of the problem lies in the difficulty of making the right decisions in the identification and development of components compatible with the requirements and meeting the needs, as well as the management of data and communication between these components.

The objective of this work is to develop a component modelization approach and to make available a manager to take the task of the system easier due to the multitude of information to consider, expressed by:

- A process for developing the system's software components and subsequently the system's architecture.
- A communication management approach between these identified software components representing the different types of users of the system.

Since the role of the components is to provide the business functionality of the system, the proposed approach is to identify the different components of the system, and keep these components throughout the software development cycle.

With the advent of UML 2.0, it has become increasingly easier to express needs and requirements such as (Karaa and al., 2016). The main addition to the UML is the new concepts.

Although originally designed as a notation for object-oriented modeling, UML has great potential for component-based systems. Components applies UML to the world of component architecture, demonstrating how it can be used to specify components, their interactions, and their integration into cohesive systems.

The goal of these approach patterns is to identify and formalize the recurrent design schemas in a context based on the UML architecture design model manipulation tool and to examine their integration into an EJB.

SOLUTIONS AND RECOMMENDATIONS (SECTION 3)

A. Design & Management Method

1. Method of Software Component

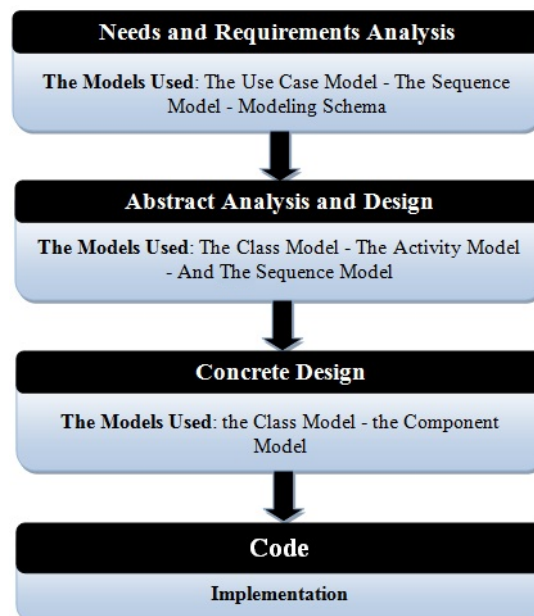
With the aim of developing information systems organizing data in an easily accessible way, or well suited to a good decision-making represented in an intelligent way towards the development of a software project, from which it is necessary to define a method of analysis and design of software components.

The data are classified and categorized, without repetition, in the information and their treatments. The study of the treatment of multiple components is generally simpler than the study of the overall treatment of a system, because each simpler component will be the subject of a separate study and simpler to study separately.

The approach according to our point of view must start from the general, and go towards the particular, that is to say a guided by the functionalities of the system; the system is divided into components where each must perform a function (possibly several, if the component is decomposed into subcomponents with separate functions). These components make it possible to organize and cut the system into manageable and exchangeable pieces of software.

The needs analysis enables the identification of strategic and methodological options for developing the functionalities of the components. The decision to break down a system into functionalities increases the number of components, and creates elements that are by default in the initial state, estimated. The decision to break down a system into functionalities increases the number of components, and creates elements that are by default in the initial state, estimated, and accompanies the activity of the actors either at the operational or decisional level. Each method or approach proposes specific steps for each phase. The work in stages as well as their phase facilitates the resolution and constitution of the project. Each of these phases describes dependencies with the other phases and ensures overall consistency and direct system matching during the execution of the approach.

Figure 2. The set of Steps and Models in the proposed Approach process



The desired functions of the modeling component expressing the essential functions of each component are plotted. The clipping here, becomes more detailed in a division into the functions of the parts.

The system is broken down using features that display all the main UML behavior models.

With the advent of UML 2.0, it has become increasingly easy to express needs and requirements such as (Karaa et al., 2016). The main addition to the UML is the new concepts. The system is broken down using features that display the main UML behavior models.

2. Management Component: Organization and Management component of Data and Information

This section deals with a knowledge management model. A good model can integrate various elements and show relationships and dependencies. But first, what are the components of a knowledge management framework.

The models display processes that simulate what actually occurs inside an organization system. The problem with the latter is that they are often hard to grasp and can only convey limited information so as not to become incomprehensible.

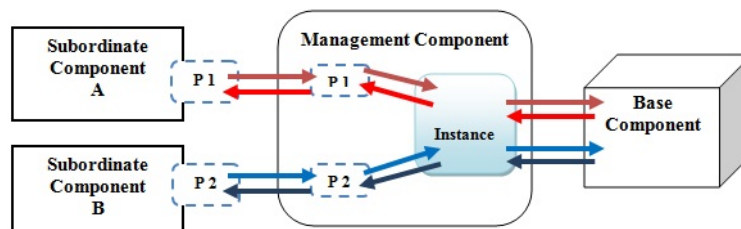
Since KM is closely dependent on other information management, project management, etc.). So a model can become very complex indeed.

Efficient management of data and knowledge are key of the success of almost any organization, but also to the successful treatment of condensed of data: with the transition to the information and the knowledge technologies, data has become both increasingly central and critical to all activities. The component processes the data through a set of activities and a library and then through the input / output functions. Component modeling is achieved by a structure consisting of the interface exposing or requesting from the rest of the application some data.

The data are classified and categorized, without repetition, in the information and their treatments. The purpose of managing information systems is for the organization of knowledge existing in information, in an easily accessible manner, or appropriately for good decision-making intelligently represented in the development of a software project.

Software systems are developed by assembling component units that provide a set of exposed services. The role of component services is the communication, the management, the organization of knowledge offered by information and the association of all the links existing between a component and another component.

Figure 3. Assembly via Management Component interacting with the Basic component and Subordinate Components



In order to decouple the management of subordinate components A and B, Management component is interposed between them. Each one invokes the Management component during the execution of the system, at which point the instance receives the call to verify the operation requested (the arrows in Figure 3 request sent and reply received).

The essential role of component-based development is the component model that defines how components can be built and how the information can be communicated and managed.

It's about defining what components the composite is composed of, how these components are interconnected and related to the composite. Composites can also include services and other attributes (Radermacher, 2013).

The steps of Management Component:

- It starts by identifying a basic component
- Then identify its subordinate components (components A and B) (Identify the components the method of the previous paragraph or Article (Rehioui, 2019))
- Interpose between the base component and the subordinate components A and B, another management component.
- The role of the management component is the management of all the links existing between a basic component and another subordinate component.
- Each process is intended to represent a component during system execution.
- Component A invokes process P1, at which point instance "I" receives the call to verify the operation requested by the process.
- The instance "I" invokes the base component, the latter checks the operation to allow or to reject the validity of the operation.
- The instance is an object with a behavior and a state.
- A subordinate component, a process is mapped.
- A set of activities is part of the whole process

The code for creating and managing all these components (and specifying how they interact with the base component) is typically defined in a Management component. This is not because it must be defined there, but because it is **logical** to separate all the information management from the system components.

Basically, it must declare a field `compManagerBase` to hold a reference to the Management component instance. It also sets up a getter to return the instance of the manager component `public ComponentManagerBase getMyComponentManager()`.

Finally, it creates a method to set the management component `compManager` by `setComponentManager(ComponentManagerBase compManager)`.

During work of the Management component, it provides the method `setprocess()` allowing activating/deactivating a process and the method `getprocess()` returning the active process. The active process is either an instance of the system component.

`Boolean ComponentIsSelected(Component comp)` Returns the selected component.

`public void setEnabled(boolean b)` enable / disable a component and there for enable/disable process, components are initially disabled by default.

Class Component extends compManagerBase {

voidComponentIsSelected(){ }

```
protected void setEnabled(boolean b){ }  
  
}
```

As a result, the architecture does not describe what a system needs to do but how it is designed to meet the specifications already built. However, the functional analysis describes the “what to do” while the architecture describes the “how to do”.

Technology (Section 4)

The software architecture expresses software components as part of information system architecture. The notion of component, as a real architectural element for the analysis and design of the information system, aims to divide and identify the system into autonomous, well-defined parts, and describes interactions with other parts.

When a piece of software or a component is created autonomous and well-defined, and describes interactions with other components, it is essential to perform several tests to ensure that the component is complete and efficient. To carry out these tests, several methods and mechanisms of code examination define how to describe abstractions and how to use them afterwards. There are generally three types of examinations: black box, white box and gray box.

The black box is represented in a basic way by displaying the inputs and outputs but by masking its structure and its internal functioning (a code), by examining only the functionalities of an application, that is to say if it does this that it is supposed to do, no matter how it does it from where we check that the end result is suitable via a “user” profile.

The white box aims to analyze a computer program whose exact structure, processes and internal functioning (the source code of the program) is known rather than its functionalities from which one has access to the complete state of the system and thus to have programming skills, in order to understand the code studied via a developer profile, and not a user.

The gray box allows access to certain information of the system state; it reveals properties of the implementation as it is an intermediate solution between the previous two is to consider gray box components. Such as giving an input (input) to a system, check that the output obtained is that expected, and verify by which process this result was obtained. In this case, a subset of the operating modes of a component would be open for modification by means of a set of predefined mechanisms.

The software component designates simple files, generally compiled code, existing in module which delivers a service, via couplings, and makes it possible to pass interactions between these modules. Each module can export or import some data in the form of variables, functions and classes later.

The era of component-oriented began to grow 30 years later by Sun EJB (Sun, 2018) OMG (OMG, 2018), Corba Component Model (CORBA, 2018), Microsoft “.NET / COM.” (Microsoft, 2018) etc. To benefit from technological advances, it is essential to adapt applications to technologies. The implementation is the culmination of the design to adapt the system as components, i.e., source code, scripts, and other elements. The main targets of the implementation are to plan the component embeddings for each recurrence, and to produce the classes in source code forms.

As a result, the architecture does not describe what a system needs to do but how it is designed to meet the specifications already built. However, the functional analysis describes the “what to do” while the architecture describes the “how to”.

The most general and the oldest is usually CORBA which suits the model of these applications. However, the heaviness and the complexity of implementing this type of application are the major disadvantages of this technology. This is why a more restrictive but powerful model has emerged: the EJB model.

EJB is this type of component that is particularly interested in business requirements through remote objects. These EJBs serve as an intermediary between windowed applications, or Web applications, and the database.

The code of a component is indeed separated into two categories; internal methods and data. One component implements “private” methods of non-externally accessible functions and “for personal evaluation”. To use them, the component must provide a means of communication with the client programs.

Specifically, an EJB component is a set of Singleton type classes and an XML file to configure it. Classes must meet a certain specification (the Stateless or Statefull EJB specification, which can be called locally, or remotely (via RMI) to execute methods), and all of which must be in a single file to form an entity.). These components, like beans, can be accessed by all types of clients: servlets, JSPs, Java clients via Java RMI (RemoteMethod Invocation). What characterizes them is the need to be in a container, called an EJB container.

It is necessary to distinguish between the categories of these components:

- **Session EJBs:** This type of bean serves one client at a time and can be seen as an extension of the client at the server.
- **Entity EJBs:** Entity beans represent data that originates or feeds the database according to a data object model. They can be accessed by multiple clients simultaneously.
 - **MessageDriven EJBs** represent a queue of messages posted by the client and processed by the server (or vice versa).

A “User class” that is transformed into an EJB by adding the @Entity annotation and having a session EJB for this entity, transforming a simple javabean into an EJB by adding the @statefull or @stateless annotation.

Some “public” or “interface” methods of externally accessible functions are intended to be called by client programs.

The interactions between the EJBs and the application server are provided by notations in order to simplify the implementation of the interfaces.

```
import java.sql.Date;
```

```
import java.util.List;
```

```
import javax.ejb.Local;
```

```
@Local
```

```
public interface LibraryManagerBeanLocal {

}
```

```
package sessions;
```

```
import java.sql.Connection;

import javax.ejb.Stateless;

import java.sql.Date;

import java.sql.DriverManager;

import java.sql.ResultSet;

import java.sql.ResultSetMetaData;

import java.sql.SQLException;

import java.sql.Statement;

import java.util.List;

@Stateless

public class LibraryManagerBean implements LibraryManagerBeanLocal {

    Connection con = null;

    Statement sta;

    ResultSet re;

    ResultSetMetaData metaData;

    .

    .

}
```

FUTURE RESEARCH DIRECTIONS

The management component approach provides a methodological idea, which leads to the development of a suite of controlled and managed components to reduce data redundancy. The future goal is to identify the typically and efficiently different parts to successfully complete the approach.

After component technology and assembly, the MDA approach (Kherraf, 2011) comes to produce models that are not only for visualization or communication, in the MDA approach, for generating and producing.

It is no coincidence that the notion of component gives a great interest in the development of complex systems. Currently, the research focuses on the complete definition of component-based information systems approach design. This type of research should lead to a real and effective method of engineering field.

CONCLUSION

The notion of software components is interested in problems such as the lack of coherence between the breakdowns of the system and the lack of traceability during the software development cycle, that are not taken into account until now.

The state of the art shows all the methods, approaches and processes aimed at undermining the complexity of the field. Moreover, the concepts are far from having a definition adopted by all, starting with the founding component. Methodological solutions based on components are in full emergence, and have shown their limits, to form simply composable entities. Software systems are developed by assembling component units that provide a set of exposed services as interfaces. The essential role of component-based development is the component model that defines how components can be built and how they can be assembled.

Currently, the research focuses on the complete definition of a component-based information systems design approach. This type of research should lead to a real and effective method of domain engineering. The goal to be achieved is to define a method of designing components in the modeling of complex systems.

The design of systems based on components supposes to mean for modeling its data management. The concept of Management Component is one of these means.

This presented approach allows the design of a software Management component by assembling a basic component and its subordinate system software components.

This approach starts with defining the different elements of the Management component whose function is to generalize all the characteristics of subordinate system components, and to forward requests from the base component. Also, it has the role of monitor and ensures the system changes state via the corresponding process and an instance of the component.

This pattern offers a good adaptation to the concept of the management and communication between the components, because it allows the system to communicate according to its priorities.

Methods are generally presented a particular vision of how to develop software, and they do not say how to program; these are processes that provide steps to drive development.

Finally, this work is aimed at software engineering researchers, who are familiar with Component-oriented Analysis, Design, and Management and want to obtain an overview of the technologies that are enabling component-based development.

REFERENCES

- ACCORD Project. (2002). *Component Assembly by Contracts in Open and Distributed Environment. Technical Report, RNTL, May 2002.* Author.
- Allier, S., Sadou, S., Sahraoui, H. A., & Fleurquin, R. (2011). From Object-Oriented Applications to Component-Oriented Applications via Component-Oriented Architecture. *WICSA, 2011*, 214–223.

Ammour, S., Beugnard, A., Matougui, S., & Traverson, B. (2004). *Software architecture and communication abstractions: an application in the field of scientific computing*. RNTL ACCORD, NOTERE2004-ACCORD.

Bali, R.K., Wickramasinghe, N., & Lehaney, B. (2009). *Knowledge management primer*. DOI: doi:10.4324/9780203894712

Boussier, J. M. (2007). Modeling behaviors in dynamic systems: Application the organization and regulation of parking and travel in Traffic Systems Urban. Software Engineering, University of La Rochelle.

Brunet, E., & Ermine, J.-L. (1994). Problem of Knowledge Management Organizations. *Information Systems Engineering*, 2(3), 263-291.

Cheesman, J., & Daniels, J. (2001). *UML Components - A Simple process for specifying Component-Based software*. Addison-Wesley.

CORBA. (2018). <http://www.omg.org/spec/CORBA/3.3/>

D'Souza, D.F., & Wills, A.C. (1998). *Objects, Components, and Frameworks with UML - The Catalysis Approach*. Addison-Wesley.

Davenport, T. H., & Prusak, L. (2000). *Working knowledge: How organizations manage what they know*. Harvard Business School Press.

EJB. (2018). <https://www.infoq.com/articles/EJB-Glossary>

Ermine, J. L., Chaillot, M., Bignon, P., Charreton, B., & Malavieille, D. (1996). MKSM, a method for knowledge management. *Advances in Knowledge Management*, 1, 288-302.

Fabresse, L., Bouraqadi, N., Dony, C., & Huchard, M. (2012). A language to bridge the gap between componentbased design and implementation. *Journal of Computer Languages, Systems & Structures*, 38(1), 29-43.

Kamal, S., Dey, N., Ashour, A. S., Ripon, S., Balas, V. E., & Kaysar, M. S. (2017). FbMapping: An automated system for monitoring facebook data. *Neural Network World*, 27(1), 27-57. doi:10.14311/NNW.2017.27.002

Kamal, S., Dey, N., Nimmy, S. F., Ripon, S. H., Ali, N. Y., Ashour, A. S., ... Shi, F. (2016). Evolutionary framework for coding area selection from cancer data. *Neural Computing & Applications*, 1-23.

Karaa, W. B. A., Ben Azzouz, Z., Singh, A., Dey, N. S., Ashour, A., & Ben Ghazala, H. (2016). Automatic builder of class diagram (ABCD): An application of UML generation from functional requirements. *Software, Practice & Experience*, 46(11), 1443-1458. doi:10.1002pe.2384

Kruchten, P. (1998). *The Rational Unified Process: An Introduction*. Editions Eyrolles.

Ludovic, L.-S. (2011). Models and tools for capitalization of knowledge in design: contribution management and knowledge engineering at Renault - DCT. Central School.

McIlro, M. D. (1968). Mass produced software components. *Proceedings NATO Conference on Software Engineering*.

Microsoft. (2018). *NET Framework Developer Center*. Author.

Mokni, A. (2015). *A formal approach to automate change management in component-based software development processes*. LGI2P.

Nimmy, S. F., Kamal, M. S., Hossain, M. I., Dey, N., Ashour, A. S., & Shi, F. (2017). Neural Skyline Filtering for Imbalance Features Classification. *International Journal of Computational Intelligence and Applications*, 16(03), 1750019. doi:10.1142/S1469026817500195

OMG. (2018). *O.M.G. Home Page*. <http://www.omg.org>

Pajić, A., & Bečejski-Vujaklija, D. (2016). Metamodel of the Artifact-Centric Approach to Event Log Extraction from ERP Systems. *International Journal of Decision Support System Technology (IJDSST)*. Advance online publication. doi:10.4018/IJDSST.2016040102

Radermacher, A., Cuccuru, A., Gérard, S., & Hamid, B. (2013). Split of Composite Components for Distributed Applications (regular paper). In *Proceedings of the Forum on Specification & Design Languages FDL 2013*. Paris, France: IEEE Computer Society.

Sosnin, P. (2016). Precedent-Oriented Approach to Conceptually Experimental Activity in Designing the Software Intensive Systems. *International Journal of Ambient Computing and Intelligence (IJACI)*. Advance online publication. doi:10.4018/IJACI.2016010104

Sun. (2018). *E.J.B. Homepage*. <http://www.oracle.com/technetwork/java/javaee/ejb/index.html>

Sylvain, C., Seriai, A., Oussalah, M., & Tazilmat, D. (2007). Extraction of Component-Based Architecture from Object-Oriented Systems. *The Third International ERCIM Symposium on Software Evolution*.

ADDITIONAL READING

Kamal, S., Dey, N., Nimmy, S. F., Ripon, S. H., Ali, N. Y., Ashour, A. S., ... Shi, F. (2016). Evolutionary framework for coding area selection from cancer data. *Neural Computing & Applications*, 1–23.

KEY TERMS AND DEFINITIONS

Base Component: It is the first common component that includes the features and common functionalities between the different components, accessible by every user, also called default component because it is the unique component that is activated when no component of the system is enabled.

Component: It is a data concept informally corresponding to a list of activities, providing the business functionality of the system.

Data: Facts and numbers which relay something specific, but which are not organized in any way and which provide no further information.

EJB Component: It's this type of component that is particularly interested in business requirements through remote objects. These EJBs serve as an intermediary between windowed applications, or web applications, and the database.

Information: when data become contextualized, calculated, its called information.

Knowledge: Knowledge is contextual information that provides an environment and framework for evaluating and making decision and information. In organizations, it often becomes implanted in documents, practices.

Management Component: Mediator component to manage all operations and to react and interact with other components. Its role is the management of all associations and links existing between a base component and the other subordinate system components, the management of the ability to communicate units.

Modelization: The process of modeling, constructing on the base of the function should give the functional modelization of the system, based on the conceptual function.

Subordinate Component: Is performed by a structure consisting of the interfaces offering or requesting some data from the base component of the application.

An Introduction to Clustering Algorithms in Big Data

5

Rajit Nair <https://orcid.org/0000-0002-4564-0920>*Jagran Lakecity University, India***Amit Bhagat***Maulana Azad National Institute of India, India*

INTRODUCTION

In Big Data, clustering is the process through which analysis is performed. Since the data is big, so it is very difficult to perform clustering approach. Big data is mainly termed as petabytes and zeta bytes of data and high computation cost is needed for the implementation of clusters. In this chapter we will show how clustering can be performed on Big Data and what are the different types of clustering approach. The challenge during clustering approach is to find observations within the time limit. Chapter also covers the possible future path for more advanced clustering algorithms. The chapter will cover single machine clustering and multiple machines clustering which also includes parallel clustering.

Today, data is increasing rapidly which actually forms a big data due to its high velocity, huge volume and different varieties of data (Shobana & Kumar, 2015). Few years before we are dealing with data collection challenges, but now in this era we are more concerned with processing this huge amount of data or big data (Nair & Bhagat, 2018a). Big data is analyzed for prediction and analysis purpose (Tsai, Lai, Chao, & Vasilakos, 2015). Scientists and researchers believe that today one of the most important topics in computer science is Big Data, whether the data collected through social networking websites, ecommerce websites or any other websites which is having accessibility to billions of users. YouTube has more than 1 billion users which is producing almost 200 hours of video in each hour and its content ID service scans over 400 years of video every day. Main reason behind storing these types of data is further used for knowledge discovery (Chu, 2013). Data mining is the method which is mainly used for knowledge discovery and clustering is the process which is used for dividing the data into groups that contain objects of similar patterns (Chu, 2013). Clustering is vastly used in many areas such as bio-informatics (Y. Q. Zhang & Rajapakse, 2008), machine learning (Kubat, 2017), pattern recognition (Neal, 2007), networking (Sucasas et al., 2016) and lot of research work is done in this area. A decade before clustering algorithm was applied on small data in order to handle their complexity and computational cost, this also increased the speed with scalability. Due to occurrence of Big data in recent years more challenges are added to perform clustering.

Big Data Challenges With its Important Characteristics

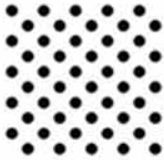
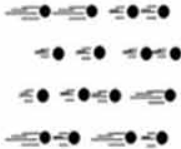
Big data is the phenomena which mainly work on 5 v's (Saporito, 2013), which is volume, velocity, variety, value and veracity, but it is not limited to 5 v's it can be extended further also. Normally Big data is unstructured, this means data cannot be easily transformed in the form of rows and columns and this

DOI: 10.4018/978-1-7998-3479-3.ch040

cannot be processed by traditional databases. So to process this Big Data, NoSql database (Lourenço, Cabral, Carreiro, Vieira, & Bernardino, 2015) are used. Let's discuss these V's one by one.

- **Volume** – This term is directly related to space occupied during data storage, whenever there is talk about big data, volume is the main aspect. Today most of the application are based on big data like data collected by Facebook which is having more than 2 billion users, data collected by traffic sensors, data collected by aero plane during transfer travel from one place to another, data collected by ecommerce website, data collected by Youtube having one billion users etc,. The amount of data collected is more than terabytes can be considered as big data but there is no limit defined to which we can say the data is big data.
- **Velocity** – This term is directly related to speed of incoming data which means when we are uploading the data at some specific server, in that case it may happen that many people are accessing that server at a time, so it experience a high speed. In case of facebook, 900 million photos are uploaded, 500 million tweets posted on twitter, more than 3.5 billion searches performed in google, 0.4 million hours videos are uploaded on Youtube. The incoming speed is just like nuclear explosion, so Big Data processing help us to deal with this high velocity of incoming data.
- **Variety** – Many of the website applications has to process and analyze different varieties of data like text, image, video, etc. and it is not an easy task to process different varieties of data. Facebook is the best example where we deal with different varieties because different users used to post textual data, images and videos at the same time, so facebook has to handle these varieties of data and later on processing should be done. To process this Big Data plays an important role.
- **Veracity** – It relates to trustworthiness of the data means the data which is collected that is reliable or not even it can be said that it is accurate or not. Veracity is very much needed during processing of big data because if we process the inaccurate data, in that case the analysis will also be inaccurate.
- **Value** – Finally the last V i.e. value which actually transforms the data in the form of value. In the complete chapter we will discuss the challenges, methods and applications of clustering in detail with respect to big data. It's mainly related to cost or profit included in the system in terms of value.

Figure 1. 5 V's of Big data

Volume	Velocity	Variety	Veradty	Value
 Stationary data It's a stationary data ranged from	 Moveable data Streaming of data that requires	 Varieties of Data Consist of structured and	 Suspicious data Uncertainty occurs due to incomplete,	 Transformation Data is converted into money for the

Applications Based on Big Data

In today's world many companies are dealing with Big data, Some of the well know industries are given below:

- **Public Administration** – Most of the governmental organization is collecting data from different areas like, Query about government projects, complaint about the government services, transportation system and many more. Government is already investing lots of money and time to process this big data (Schmid, 2017).
- **Healthcare and Social care** – Healthcare industry is one of the growing industry in the area of Big data because the data i.e. collected is mostly in the form of images and videos, it can be like patients MRI images or videos related to operations of the patients. This way huge amount of data is collected and it has to be processed in very less time. There can be other textual data also which will be in the form of like medical report of the patient, complaint regarding to service of hospitals and many more (Rahman, Slepian, & Mitra, 2016).
- **Transport and logistics** – Today most of the transportation and logistics are GPS enabled though which one can easily track the location of the objects. But at the same time these objects are collecting large amount of data and it depends that they may be processed or not. If any issue occurs like accident, then definitely it will be analyzed otherwise it depends on the conditions (Nair & Bhagat, 2018b).
- **Retail and Trade** – Considering retail and trade, e-commerce is the best example and it is used by many people for selling and buying. Large amount of data is collected during the process done in e-commerce like surfing for product, purchasing, selling, reviews for the products, product details, customer information, etc (Aktas & Meng, 2017).
- **Geographic information system (GIS)** – GIS industry has its own nomenclature when dealing with big data and named it as Spatial Big Data (SBD). They are also facing the problem of huge volume, variety and update frequency, which cannot be handled by usual spatial computing techniques, architecture, methodologies and software solution. Nowadays big data processing techniques are used for GIS (Wu, Zhang, & Gong, 2014).
- **Media and Entertainment** – Another broad area of big data, in this huge amount of data is collected in the form of news paper content, audios, videos and many more. Polling and reviews are also stored in the form of data. A popular websites IMDB that contains all the details of all the movies which are released today anywhere in the world. It contains the ratings, reviews and all the information related to movie (Mahrt & Scharkow, 2013).
- **Application and Services** – Most of the application and services based company are collecting huge amount of data in the form of user's details, reviews, request and the operations done by user (Dobre & Xhafa, 2014).

BACKGROUND

Before discussing the specific techniques of clustering, this section will provide some necessary background, first we will define the cluster analysis after that illustrate the problems during cluster analysis and describe its relationship with other techniques that group the data.

Cluster Analysis

In order to extract data from Big Data, cluster analysis is very much needed. This is the place Cluster Analysis systems can characterize and group the objects based on attributes. Cluster Analysis looks to apply a typical way to deal with collected objects on the basis of their relationship to each other and by accessing the structure generated by the user selected characteristic. It is a nonexclusive name for a gathering of multivariate data analysis techniques that use a cluster variate, a scientific representation that compares the similarity of objects. Cluster analysis is widely used in pattern recognition, by recognizing data patterns analyst can reduce the number of observation which he has to make and can predict some hypothesis about the structure of the problem which has to be tested.

Clustering analysis is especially helpful in situations identified with selecting between various business choices, where the choice relies upon the similarity (or contrast between) groups. It very well may be applied to portraying or delivering a scientific classification of objects, for rearranging information where there are numerous perceptions, or for distinguishing connections between items. Regular business models incorporate client division, utility or power utilization and client conduct profiling. Advanced clustering strategies are likewise utilized in more mechanically advanced applications like picture, number plate or facial acknowledgment frameworks, however the standard of classification dependent on a known or assumed cluster variate continues as before. Now in the next section we will see the clustering approaches used in big data.

FOCUS OF THE ARTICLE

The focus of this article is based on clustering algorithms in big data. Let us discuss these algorithms in brief

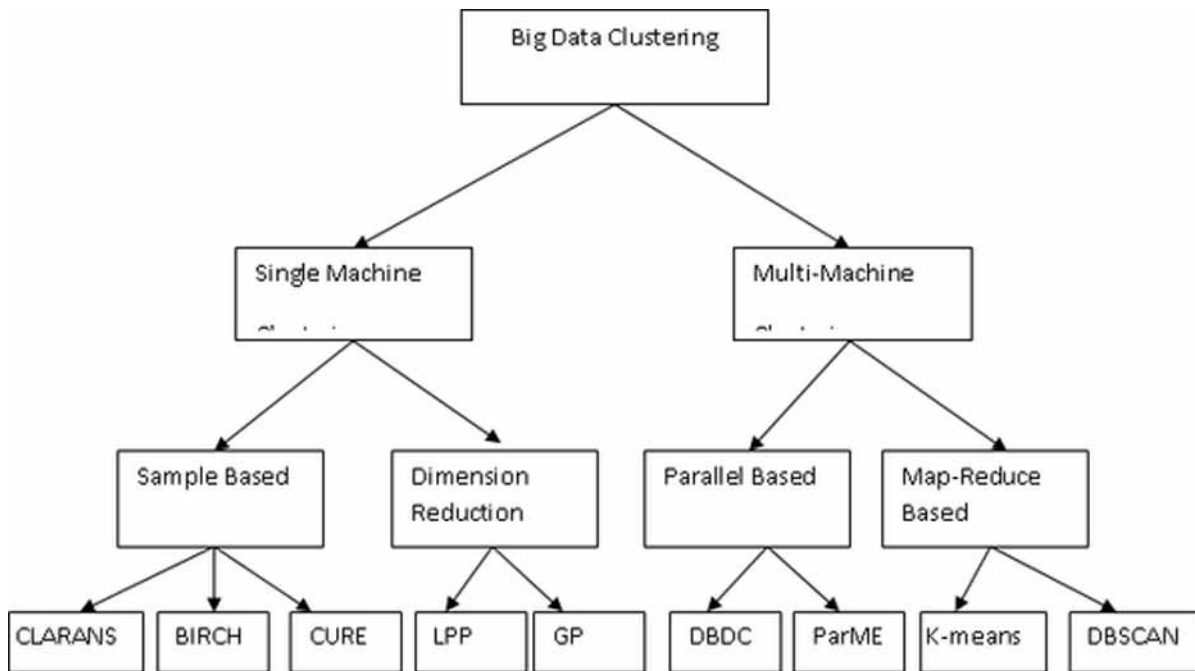
Clustering

Clustering is mainly used in data mining for analysis purpose, in which similar elements are grouped together. Machine learning is classified into two main categories one is supervised and the other one is unsupervised (Kotsiantis, Zaharakis, & Pintelas, 2006). Clustering comes under the category of unsupervised machine learning which is used as a statistical analysis in many fields. Clustering analysis is used to gain some useful insights from our data by identifying the group in which the particular data belongs. Generally big data clustering is classified into two categories one is single machine clustering and the other one is multiple machine clustering. Nowadays multiple machine clustering has gained a lot of popularity due to its flexibility in extension and it offers faster response to users during processing. In the below section we have shown the different techniques based on these clustering categories. Sample based techniques (Daszykowski & Walczak, 2010) and dimension based techniques (Bouveyron & Brunet-Saumard, 2014) are based on single machine clustering, parallel (W. Y. Chen, Song, Bai, Lin, & Chang, 2011) and map-reduce (Rasheed & Rangwala, 2013) are based on multiple machine clustering.

Let's discuss the figure in detail,

1. **Single Machine clustering** - This clustering is perform on single system and easy to implement. This clustering is fruitful when fewer amounts of data is available and handling of data can be done by single system.

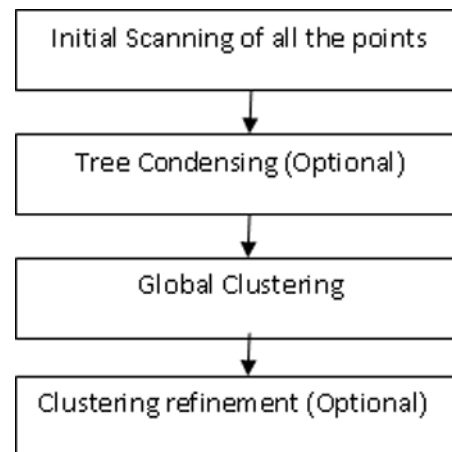
Figure 2. Classification of Clustering Algorithm



- Sampling Based Techniques – These algorithms work on sample from the large dataset that means clustering will be performed on sample and then generalize to complete dataset. This was the first method which attempts to improve the speed and scalability and it is dealing with exponential search space. Computation speed is high due to less amount of space required during clustering operation and complexity also decreases due to small amount of data. Some of the algorithms based on sampling technique are as follows:
 - (i) Clustering Large Applications based on Randomized Sampling (CLARANS) – Before CLARANS there are some more algorithms like PAM (Partition Around Medoids) (Van der Laan, Pollard, & Bryan, 2003) are introduced but these are not sufficient for large data set clustering. So to perform large dataset clustering CLARANS is introduced that decreases the overall quadratic complexity and time requirements into linear in total number of objects. PAM used to calculate the entire pair-wise dissimilarity matrix between objects and store it in central memory but in case of CLARANS there is no need to compute dissimilarity matrix. It is considered as a graph searching problem in which each node is considered as possible clustering solution and criteria for combining two nodes is that they should differ in 1 out of k medoids. CLARANS aims to find a local optimal solution by investigating the entire graph and during iteration this only checks a sample of neighbors of the current node. Dynamic sampling is done at iteration of the search procedure due to which CLARANS has improved performance over CLARA and PAM.
 - (ii) Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) – It's an unsupervised learning algorithm used to perform hierarchical clustering particularly over large datasets. The main advantage of BIRCH is its ability to form a best quality cluster for the given set of points through incremental and dynamic approach. If the size of the

processing data is larger than the memory size, in that case computational cost will also get increases in the form I/O devices. This algorithm uses its own data structure called as clustering feature (CF) or clustering tree. Clustering feature is a triplet that contains the number of points in the cluster, the linear sum of the data points and the square sum of the data points in the cluster. It assumes that not always all points are equally important for clustering and at the same time it is not needed that all the points has to be loaded in main memory. If clustering feature satisfies the additive property then two clusters can be merged together, the CF will be sum of the CF's of the other two clusters. This algorithms works in two steps: first it scans the data points and then builds a memory tree according to those points. Secondly cluster algorithm is used to cluster the leaf nodes. According to (T. Zhang, Ramakrishnan, & Livny, 1996) BIRCH performs better than CLARANS and it is adaptive for handling outliers. Given below figure 3 shows how BIRCH algorithm works.

Figure 3. Steps in Birch algorithm

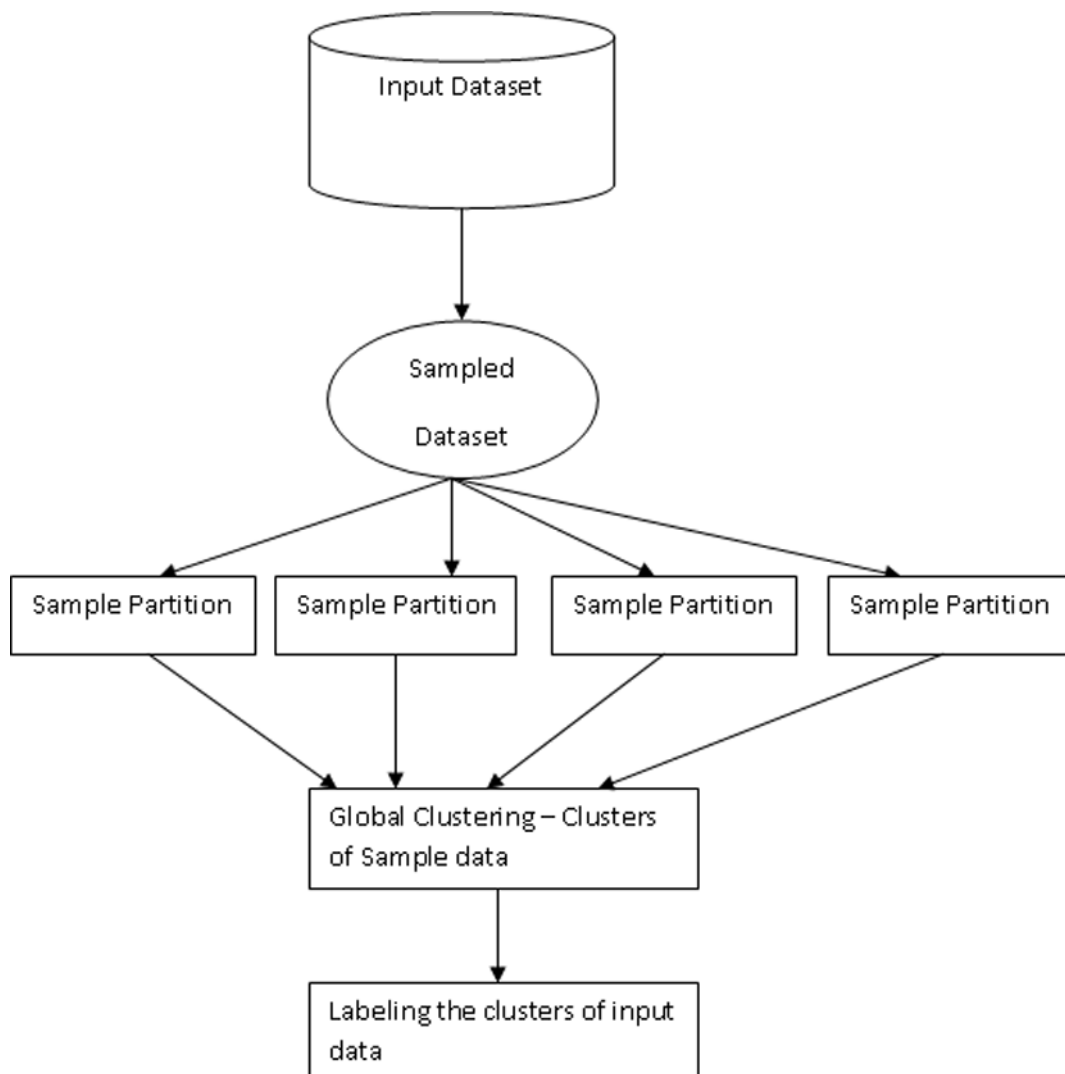


- (iii) CURE: The above two discussed algorithms focus on the points and later on they are considered as clusters, these algorithms performs well only when the clusters have spherical shape. While dealing with real time applications most of the clusters form complex shapes. To handle these complex shape clusters, CURE (Clustering by using representatives) algorithm has been evolved that uses a set of well scattered data points to represent a cluster. CURE is also a hierarchical clustering algorithm that considers single data point as cluster and continuously combines two existing clusters until it reaches to precisely k remaining clusters. The selection of clusters for merging are based on the calculated minimum distances between the possible pair of representative points from the existing cluster. CURE uses two data structures for efficient search one is heap and the other one is k-d tree, heap is used to track the distance for each existing cluster to its neighboring cluster and k-d is used to store all the representative points contained in the cluster.

This algorithm is mostly used for large databases, it solve the problems by applying two main methods, sampling and partitioning. Sampling method is used to speed up the computation during cluster formation. It draws the sample from the large dataset and applies the procedures on these sample data.

To decide the sample size Chernoff bound is used in original work. After sampling, still it is observed that data size is very big and it cannot be processed by simple clustering method, in that case partition methods will be applied to accelerate the algorithm. Suppose there is X as the original dataset, X^1 as the sample data that is divide into y partitions and in each partition it applies the partial hierarchical clustering until it achieves a predefined number of clusters or the distance between two clusters meet the certain threshold value so that they can combine or merged. Now clustering runs and passes on all partial clusters from all the y partitions. Performing all these steps, at the final stage, all the non-sample data points will assign to nearest clusters. Results show that CURE algorithm (Guha, Rastogi, & Shim, 2001) have lower execution time than BIRCH and it also maintain the privilege of robustness during handling outliers by compressing the representative points to the centroid of the cluster with a constant factor.

Figure 4. Steps in CURE algorithm



- **Dimension Reduction** – Normally big datasets are high dimensional datasets, so when we are dealing with high dimensional datasets in that case consumption of time and complexity gets increases (Donoho, 2000). It is very much needed that dimensions should be reduced so that speed and performance of clustering gets increased. The above explained sampling method only reduces the dataset sizes but the dimensions will remains same. To reduce the dimensions there are mainly approaches two approaches, random projection and local projection.
 - (i) **Random Projection (Locality-Preserving Projection)** – Transformation in dataset is done from d dimensional to lower dimensional space, during this pair wise distance is preserved. Most of the clustering algorithm is distance based and it is assumed that clustering formed from the resulted projection is approximately equal to the original clustering space. Linear transformation of original matrix is performed for random projection, let M be an original matrix and R be a rotation matrix of $d * t$ (t is dimensions, $t < d$) and its element $R(i, j)$ be independent random variables is transformed to $M' = M.R$ that contains x dimensions which is very less than t . M' is projection of matrix M with x dimensions. Although in case of random projection construction of rotation matrix is different. There can be many more methods for performing random projection. After transforming projected matrix into lower dimensional space, clustering can be accomplished (Shikkenawis & Mitra, 2016).
 - (ii) **Global Projection** – In this projection the main aim is to find the data point which is to be projected as close as to the original data point which is unlikely to random projection that works on the concept of maintaining the same pairwise distance between the points for the projected space as that of original space. If the original matrix is M and M' be the projected matrix then according to global projection the aim is to find $|M' - M|$.
- 2. **Multiple Machine Clustering** – Above discussed sampling and dimensional reduction algorithms are based on single machine clustering that also increases the scalability and speed of the algorithm. But nowadays what we have observed that the data which is collected is very huge as compared to data collected 10 years before, they are in terabytes and petabytes, which needs faster memory and processors and this cannot be handle by single processor and memory. Due to this high storage processing problem, there is requirement of multiple processor and need of algorithms that can run on multiple processors. This technique allows huge data into smaller pieces which can be loaded into different machines and utilize their processing power for computation of individual result. To perform this operation multiple machine clustering algorithm is classified into two categories:
 - **Parallel Clustering Algorithm (Un-automated distributing)**
 - **Map reduce Clustering Algorithm (Automated distributing)**

Above figure shows hoe clustering is done through multiple machines, here data D is split into $D1$, $D2$, $D3$ and $D4$ and they are processed individually after that we combine the individual results R , so that they can form the cluster.

In case of parallel clustering, load distribution is of the most important process, because other than parallel processing challenges you have to distribute your load to different machines depend on their capability which make complicated and time consuming. Comparing to parallel clustering approach, map reduce approach is better because it allows programmers to work comfortably and reduce the burden of programmers related to load distribution, unnecessary networking problems, fault tolerance etc. Due to these automatic features the parallelism is achieved easily in very less time. Parallel and distributed

Figure 5. Multiple machine clustering techniques

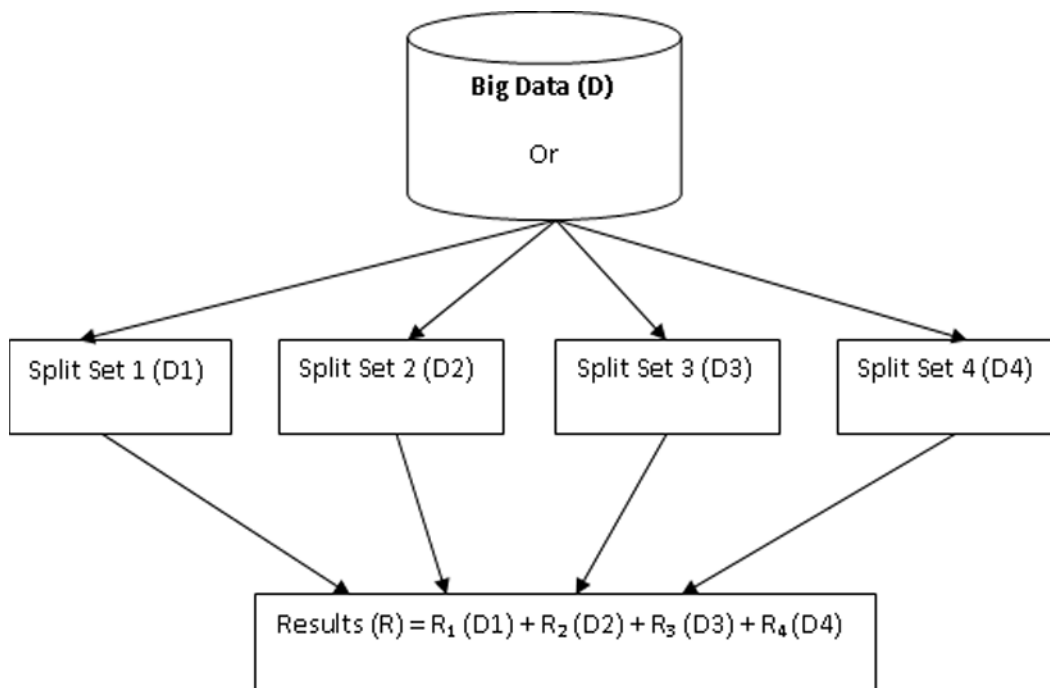
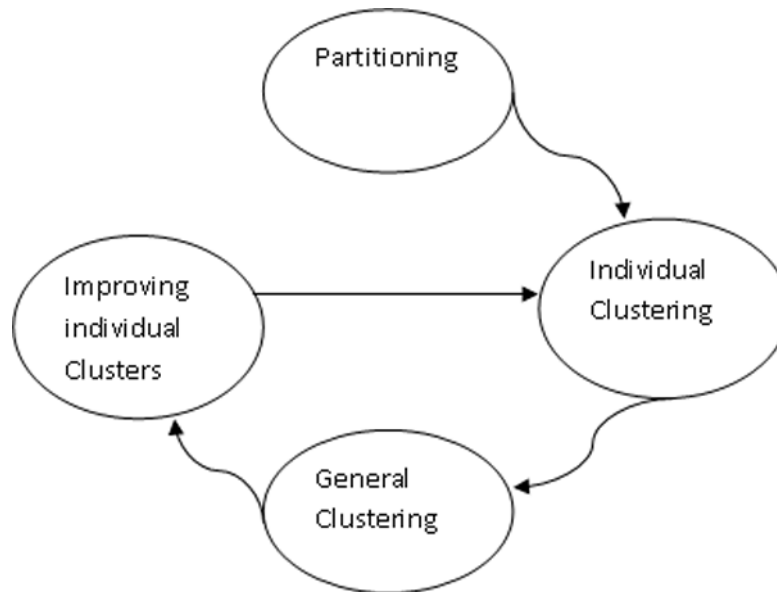


Figure 6. Lifecycle of multiple machine clustering algorithms



algorithm (Dhillon & Modha, 2007) have some of the common characteristics like, dividing the large dataset into small size data set and processing these small datasets in multiple machines individually to form a small clusters. These small clusters are combined together which later on try to form a bigger cluster. The bigger challenges that they face are minimizing the data traffic and less accuracy as compared to serial processing. Given below figure will represent the cluster formation during parallel strategies.

Lesser accuracy might be achieved due to following two reasons, one is applying different algorithms in different machines and second one can be applying the same algorithm in multiple machines. In some case divided data can change the final observation of clustering.

- Parallel Clustering - Comparing to serial clustering, parallel clustering is difficult to implement but it provides speed and scaling of clustering algorithms. Now we will see some of the well known parallel clustering algorithm (W. Y. Chen et al., 2011).
 - (i) DBDC (Density Based Distributed Clustering) – The main objective of DBDC is to form clusters of arbitrary shapes. High density of points is maintained inside the cluster than the outside cluster points but the noise region is more inside as compared to outside of the cluster. This algorithm almost follows the steps shown in above figure 6. At the individual clustering level they use some defined algorithms and at general clustering level they proposed DBSCAN i.e. a single machine based clustering algorithm for computing the final result. It is sure that even DBDC maintain the same clustering quality as that of DBSCAN but when it comes to speed it run almost 30 times faster than the serial approach (Januzaj, Kriegel, & Pfeifle, 2004).
 - (ii) ParMETIS – This is another parallel processing algorithm based on METIS for distributed system. METIS is a multilevel partitioning algorithm. Mainly there are three phases in ParMETIS that are coarsening, partitioning and uncoarsening. According to first phase maximal matching on the actual graph is performed after that similar vertices are combined together to form reduced graph. This process until we reach to some minimum number of vertices. Now the second phase in which multilevel recursive bisection algorithm is applied on the k-way partitioning of the graph obtained from the first phase i.e. coarsened. In the last phase i.e. uncoarsening, the greedy retirement algorithm is applied to project the partitioning of the second phase to the original graph. Although this above paragraph discussion is about METIS but these given three phases are used in ParMETIS also and this algorithms is considered as distributed approach of METIS. ParMETIS uses graph for formation of clusters, it does not follow the general cycle for distributed and parallel clustering which is discussed above. This algorithm was mainly designed for clustering and partitioning graphs that occur in various computational science applications and it not much suited for social networking applications which consist of millions of vertices and edges in the graph due to its limited amount of memory (Sui, Nguyen, Burtscher, & Pingali, 2011).

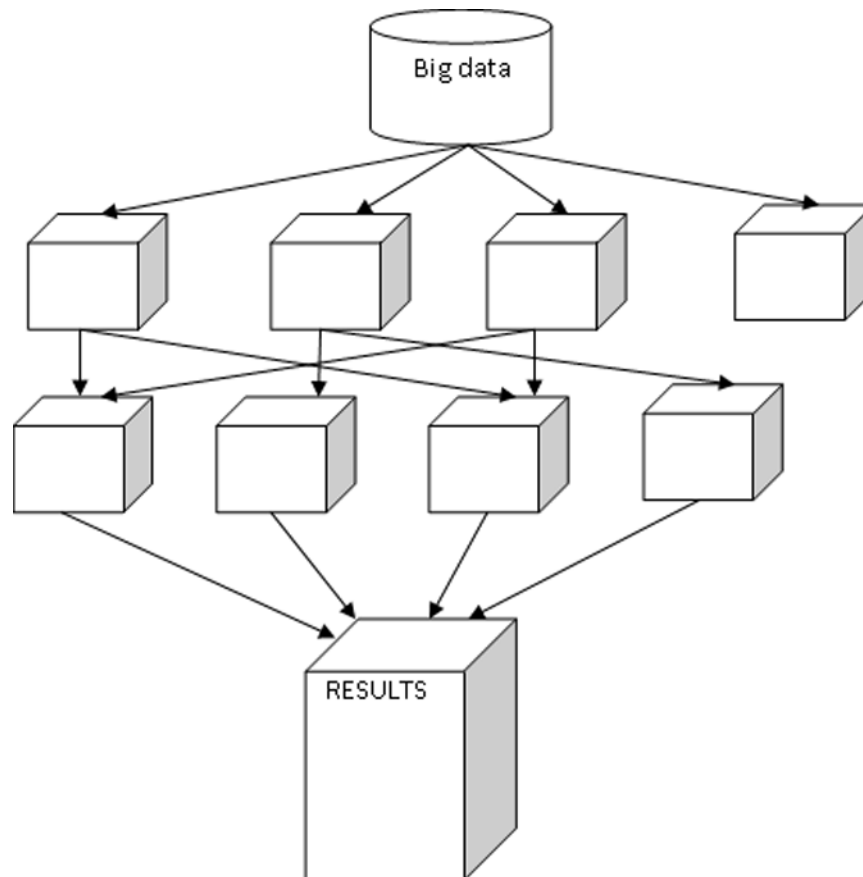
ParMETIS method performs coarsening and uncoarsening operations that are considered to be graph operation and this makes them different from other clustering approaches. At initial stage each processor receives an equal number of vertices, in the next phase i.e. coarsening in which graph coloring process has been done and then global graph is computed by incrementally matching the vertices of the same color one at a time. Now at the partitioning phase all the coarsened graph is broadcasted to all the machines, after receiving the graph each graph will performs recursive bisection by exploring the single path of the recursive bisection tree. In the last phase i.e. uncoarsening, same color vertices are used for movement, subset of these vertices leads us to reduction in elements. During movement process not all the vertices move, only the selected vertices through ParMETIS will move and this helps in the reduction of communication cost. It has been observed that running time of PerMETIS is 15-35 times faster than serial algorithm when computation is done using 128 processors.

Clustering Approach Through Graphics Processing Unit (GPU)

In recent times parallel processing has gained a lot of importance due to their high computation speed, to perform this GPU (Graphics Processing Unit) is widely used. GPU has more processing power than CPU (Buluç, Gilbert, & Budak, 2010), this is the reason nowadays all the applications are using GPU whether it is Machine learning, deep learning, Image processing, data analytics or blockchain (Morishima & Matsutani, 2018), almost all of them are using GPU. Clustering algorithm is also proposed using GPU i.e. based on DBSCAN named as G-DBSCAN (Andrade et al., 2013), it's a GPU accelerated parallel algorithm for density-based clustering algorithm. This algorithm use graph based data indexing to add flexibility and this increase parallelization process. G-DBSCAN algorithm works in two steps and both the steps has been parallelized. In the first step construction of graph is done in which each object represents a node and an edge is created between nodes on the basis of distance or some predefined threshold value. After the construction of graph, the second step is identification of clusters, to perform this Breadth First Search (BFS) is used to traverse the graph that has been created in the previous step. It has found that in comparison to serial implementation, G-DBSCAN is almost more than 110 times faster.

- **Map Reduce** – Above discussed parallel clustering algorithm provides us the scalability and increase in speed, despite all these advantages there is big challenge during dealing with memory and distribution of objects to different processors. To overcome from these challenges Google introduced the framework known as MapReduce (Dayalan, 2018). Hadoop uses MapReduce for processing BigData. The improvement by using MaReduce is shown by three factors, which are Speed-up, Scale-up and Size-up.

Figure 7. Map-Reduce framework



- Speed-up: This factor relates to the ratio of execution time when dataset size remains constant and number of machines gets increased.
- Scale-up: When machine size gets increase in that case if job size also get increase with same execution time.
- Size-up: If the number of machines remain constant, in that case during increase in data size will increase the execution or running time linearly.
 - (i) Parallel K-means clustering (MapReduce based K-means) – This clustering is distributed version of K-means clustering. K-means is one of the best clustering algorithm used in serial clustering system, which divides the whole dataset n into k clusters ($n \gg k$). Initially it assumes all the n points are independent of each other and select K random points as a cluster and calculate the nearest points from each cluster by computing the Euclidean distance between the points and the cluster. After forming the k clusters of n points, now it updates the centre of each cluster with mean of the instances. These steps will be applied repeatedly till it achieves the correct form of cluster.

PK means also work on the same concept but it runs on multiple machines with slight changes, it distributes computation to multiple machine by using MapReduce approach to enhance the computation speed and scaling of the process. As discussed above MapReduce is a two-step process, so in the first phase individual cluster will be formed using mapper method and in the second phase the entire individual clusters will be merge to form general cluster using reducer approach. It also exhibit linear speed-up and size-up property with good scaling (Zhao, Ma, & He, 2009).

- (ii) MR-DBSCAN: This is another recent map reduce approach algorithm based on DBSCAN, that overcome the limitations exist in parallel DBSCAN algorithm. Main problems occur in the parallel DBSCAN are:
 - (i) Balancing during load distribution to multiple systems.
 - (ii) Some of the critical procedures are not enough parallelized due to which there is problem in algorithm scalability.
 - (iii) Less portability is another problem occurs due to their architecture and design.

MR-DBSCAN is the other partitioning algorithm based on DBSCAN which achieves fully parallelism with cost efficient scalable approach. Some of the experiments have shown how MR-DBSCAN works efficiently on the large dataset as compared to other algorithms.

MapReduce based on GPU - GPU's are much more efficient than CPU, using MapReduce with CPU provide efficient framework for distributed computing but performance can be improved in terms of speed and scalability by using GPU (He, Fang, & Luo, 2008). Another algorithm is also proposed based on the concept of MapReduce framework that can run on multiple GPU's named it as GPMR (Stuart & Owens, 2011). Still it is in the development stage but many researchers claimed that due to rapid increase in data, this type of algorithm will be useful for more fast and scalable clustering.

FUTURE RESEARCH DIRECTIONS

This section will discover some of the areas related to clustering in which future research can be done

- Clustering techniques have done good job for clustering numeric datasets by fundamentally improving the base clustering algorithms but still lot of work has to be done for the mixed dataset (Vega-Pons & Ruiz-Shulcloper, 2011).
- Different mixed datasets in application areas, for example, medicinal and socio-financial matters contain uncertain information due to improper data acquisition methods or inborn issues in the information procurement. Clustering uncertain mixed data is an important future research in different domains.
- Some of the researchers have developed methods which can convert the mixed dataset to pure numeric dataset (Wei, Chow, & Chan, 2015). This is again the difficult problem of mixed data clustering and it has been observed that during conversion there is always a loss of information, so this has become the aspect for future research.
- It has been noticed that real world mixed data is imperfect, one of the reason for this imperfection is missing values. However some of the methods are developed which can compute mixed values for the mixed data and can perform clustering (Audigier, Husson, & Josse, 2016). The other approach which is developed to handle these missing data is clustering algorithm through their objective function (Hunt & Jorgensen, 2003). Although these approaches are not enough capable to handle the real world problems, so these missing values are creating the scope for the research.
- Another outcome of big data is guaranteeing the scalability of clustering algorithm to make them helpful in real world situations. Parallelization of mixed data clustering algorithm is a feasible methodology to enable them to scale with expanding information size and keep up linear time complexity (particularly for partitional clustering) (Subramaniaswamy, Vijayakumar, Logesh, & Indragandhi, 2015). Dynamic research around there is required to keep the field in synchronization with huge information challenges. So also, growing quick and precise web based clustering calculations to deal with huge streams of mixed information expects regard for location inadequacies. These incorporate low grouping quality, assessment of new ideas and idea float in the fundamental information, troubles in deciding cluster centers and poor ability to manage anomalies (J. Y. Chen & He, 2016).
- Another rising region of research is formation of clusters with deep learning algorithms (Min et al., 2018). The goal (loss) of deep learning clustering methods is to find deep networking loss and clustering loss. Consequently, these techniques differ as per the network architecture or on the types of clustering technique, (for example, partitional or hierarchical). However these methods are mainly applied for numeric datasets so there is a great opportunity to work with mixed data clustering with deep learning algorithms.
- It become complex when there is increment in space and size, most of the machine learning algorithms lose their interpretability and might be treated as a black box. Even mixed data clustering are not exception. Clustering models that are simple to disclose is alluring to professionals, for example, clinicians, business experts, geologists, and researcher. Interpretable models can help them in making educated choices. Tragically, just a couple of specialists have investigated this zone of creating interpretable mixed data information clustering strategies to address basic parts of the models. There has been late research in the field of highlight determination for mixed information (X. Zhang, Mei, Chen, & Li, 2016); in any case, consolidating such outcomes with clustering has not been tremendously investigated. Selecting a subset of relevant features can possibly upgrade the interpretability of clustering algorithms too.
- Spectral clustering produces great outcomes and doesn't require solid assumptions about the insights of the clusters. Spectral clustering has been utilized to group mixed datasets. The similarity

matrix is the initial step of Spectral clustering. During spectral clustering for mixed datasets builds up its very own similarity matrix (Niu et al., 2016).

- Subspace clustering is a feasible way to deal with clustering of big data with high-dimensional mixed data, however the huge information issue in itself is testing. The expansion of other subspace clustering approaches, for model, network based techniques for mixed datasets is vital to improvement of clustering algorithm for high dimensional mixed datasets (L., E., & H., 2004). In subspace clustering, an information point can have a place with more than one group and subspaces. Research on adjusting other subspace clustering approaches that have been produced for numeric datasets, for example, correlation clustering, ought to be stretched out to mixed data clustering (Immerlica & Wirth, 2008). Specifically, utilizing the connection between's numeric and categorical features to make subspaces are an inventive research zone.
- Integration of space information into clustering is a significant research region as this can improve the clustering accuracy and interpretation. Constrained clustering is a way to deal with taking care of these types of issues. Constrained clustering for iterative partitional clustering techniques has been proposed for mixed datasets (Wagstaff, Cardie, Rogers, & Schrödl, 2001). Although, there has been no research in the area of constrained clustering with other clustering algorithm like hierarchical and density-based clustering. With the availability of large scale information, there is a need to develop clustering algorithm for mixed data that can use this learning to make accurate and interpretable clusters.
- Many of the clustering algorithms require client side parameter, that's why final clustering results are based on these parameters, which incorporate number of desired clusters and initial clusters for iterative partitional clustering algorithm and the model determination for model-based clustering. A few endeavors have been made to create parameter free clustering algorithms for mixed datasets (Böhm et al., 2010). Still the research in this area is open.

CONCLUSION

It is well known that data will get increased day by day, this indicates that complexity will also get increases. So it is very much needed that there must be technology upliftment in this area which can minimize the complexity. It will always be a challenge to maintain scalability during clustering due to this rapid increase in data. In this chapter we have discussed about the clustering approaches used for data processing which includes serial clustering, distributed and parallel clustering and GPU based clustering. In the future it might be happen that integration of the above discussed algorithm will make a new clustering algorithm to gain a speed and scalability.

REFERENCES

- Aktas, E., & Meng, Y. (2017). *An Exploration of Big Data Practices in Retail Sector*. Logistics. doi:10.3390/logistics1020012
- Andrade, G., Ramos, G., Madeira, D., Sachetto, R., Ferreira, R., & Rocha, L. (2013). G-DBSCAN: A GPU accelerated algorithm for density-based clustering. *Procedia Computer Science*. doi:10.1016/j.procs.2013.05.200

- Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1), 5–26. Advance online publication. doi:10.1007/11634-014-0195-1
- Böhm, C., Goebel, S., Oswald, A., Plant, C., Plavinski, M., & Wackersreuther, B. (2010). Integrative parameter-free clustering of data with mixed type attributes. *Lecture Notes in Computer Science*. doi:10.1007/978-3-642-13657-3_7
- Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71, 52–78. Advance online publication. doi:10.1016/j.csda.2012.12.008
- Buluç, A., Gilbert, J. R., & Budak, C. (2010). Solving path problems on the GPU. *Parallel Computing*. Advance online publication. doi:10.1016/j.parco.2009.12.002
- Chen, J. Y., & He, H. H. (2016). A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data. *Information Sciences*, 345, 271–293. Advance online publication. doi:10.1016/j.ins.2016.01.071
- Chen, W. Y., Song, Y., Bai, H., Lin, C. J., & Chang, E. Y. (2011). Parallel spectral clustering in distributed systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3), 568–586. Advance online publication. doi:10.1109/TPAMI.2010.88 PMID:20421667
- Chu, W. W. (2013). *Erratum: Data Mining and Knowledge Discovery for Big Data*. doi:10.1007/978-3-642-40837-3_10
- Daszykowski, M., & Walczak, B. (2010). Density-Based Clustering Methods. *Comprehensive Chemometrics*. doi:10.1016/B978-044452701-1.00067-3
- Dayalan, M. (2018). MapReduce: Simplified Data Processing on Large Cluster. *International Journal of Research and Engineering*. doi:10.21276/ijre.2018.5.5.4
- Dhillon, I. S., & Modha, D. S. (2007). *A Data-Clustering Algorithm on Distributed Memory Multiprocessors*. doi:10.1007/3-540-46502-2_13
- Dobre, C., & Xhafa, F. (2014). Intelligent services for Big data science. *Future Generation Computer Systems*, 37, 267–281. Advance online publication. doi:10.1016/j.future.2013.07.014
- Donoho, D. L. (2000). High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. *Statistics*.
- Guha, S., Rastogi, R., & Shim, K. (2001). CURE: An efficient clustering algorithm for large databases. *Information Systems*, 26(1), 35–58. Advance online publication. doi:10.1016/S0306-4379(01)00008-4
- He, B., Fang, W., & Luo, Q. (2008). Mars: a MapReduce framework on graphics processors. *Proceedings of the 17th* 10.1145/1454115.1454152
- Hunt, L., & Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis*, 41(3-4), 429–440. Advance online publication. doi:10.1016/S0167-9473(02)00190-1

Immorlica, N., & Wirth, A. (2008). Correlation clustering. In *Constrained Clustering*. Advances in Algorithms, Theory, and Applications.

Januzaj, E., Kriegel, H.-P., & Pfeifle, M. (2004). DBDC: Density Based Distributed Clustering. *Advances in Database Technology - EDBT 2004*. doi:10.1007/978-3-540-24741-8_7

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190. Advance online publication. doi:10.1007/10462-007-9052-3

Kubat, M. (2017). *An Introduction to Machine Learning*. *An Introduction to Machine Learning*. doi:10.1007/978-3-319-63913-0

Lourenço, J. R., Cabral, B., Carreiro, P., Vieira, M., & Bernardino, J. (2015). Choosing the right NoSQL database for the job: A quality attribute evaluation. *Journal of Big Data*, 2(1), 18. Advance online publication. doi:10.1186/40537-015-0025-0

Mahrt, M., & Scharkow, M. (2013). The Value of Big Data in Digital Media Research. *Journal of Broadcasting & Electronic Media*, 57(1), 20–33. Advance online publication. doi:10.1080/08838151.2012.761700

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture. *IEEE Access: Practical Innovations, Open Solutions*, 6, 39501–39514. Advance online publication. doi:10.1109/ACCESS.2018.2855437

Morishima, S., & Matsutani, H. (2018). Accelerating Blockchain Search of Full Nodes Using GPUs. *Proceedings - 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018*. 10.1109/PDP2018.2018.00041

Nair, R., & Bhagat, A. (2018a). *A Life Cycle on Processing Large Dataset - LCPL Rajit Nair*. Academic Press.

Nair, R., & Bhagat, A. (2018b). *An Application of Big Data Analytics in Road Transportation*. doi:10.4018/978-1-5225-3870-7.ch003

Neal, R. M. (2007). Pattern Recognition and Machine Learning. *Technometrics*, 49(3), 366. Advance online publication. doi:10.1198/tech.2007.s518

Niu, K., Niu, Z., Zhao, X., Wang, C., Kang, K., & Ye, M. (2016). A Coupled User Clustering Algorithm for Web-based Learning Systems. *Proceedings of the 9th International Conference on Educational Data Mining*.

Rahman, F., Slepian, M., & Mitra, A. (2016). A novel big-data processing framework for healthcare applications: Big-data-healthcare-in-a-box. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*. 10.1109/BigData.2016.7841018

Rasheed, Z., & Rangwala, H. (2013). A map-reduce framework for clustering metagenomes. *Proceedings - IEEE 27th International Parallel and Distributed Processing Symposium Workshops and PhD Forum, IPDPSW 2013*. 10.1109/IPDPSW.2013.100

L., P., E., H., & H., L. (2004). Subspace Clustering for High Dimensional Data: A Review. *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*.

- Saporito, P. (2013). The 5 V's of Big Data. *Best's Review*.
- Schmid, A. (2017). *BigData-PublicControlling Fundamental changes in Public Management*. Public Policy and Administration., doi:10.13165/VPA-17-16-2-11
- Shikkenawis, G., & Mitra, S. K. (2016). On some variants of locality preserving projection. *Neurocomputing*, 173, 196–211. Advance online publication. doi:10.1016/j.neucom.2015.01.100
- Shobana, V., & Kumar, N. (2015). Big data - A review. *International Journal of Applied Engineering Research*.
- Stuart, J. A., & Owens, J. D. (2011). Multi-GPU MapReduce on GPU clusters. *Proceedings - 25th IEEE International Parallel and Distributed Processing Symposium, IPDPS 2011*. 10.1109/IPDPS.2011.102
- Subramaniaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Unstructured data analysis on big data using map reduce. *Procedia Computer Science*. doi:10.1016/j.procs.2015.04.015
- Sucasas, V., Radwan, A., Marques, H., Rodriguez, J., Vahid, S., & Tafazolli, R. (2016). A survey on clustering techniques for cooperative wireless networks. *Ad Hoc Networks*, 47, 53–81. Advance online publication. doi:10.1016/j.adhoc.2016.04.008
- Sui, X., Nguyen, D., Burtscher, M., & Pingali, K. (2011). Parallel graph partitioning on multicore architectures. *Lecture Notes in Computer Science*. doi:10.1007/978-3-642-19595-2_17
- Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: A survey. *Journal of Big Data*, 2(1), 21. Advance online publication. doi:10.1186/40537-015-0030-3 PMID:26191487
- Van der Laan, M. J., Pollard, K. S., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8), 575–584. Advance online publication. doi:10.1080/0094965031000136012
- Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03), 337–372. Advance online publication. doi:10.1142/S0218001411008683
- Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained K-means Clustering with Background Knowledge. *International Conference on Machine Learning ICML*.
- Wei, M., Chow, T. W. S., & Chan, R. H. M. (2015). Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. *Entropy (Basel, Switzerland)*, 17(3), 1535–1548. Advance online publication. doi:10.3390/e17031535
- Wu, H., Zhang, T., & Gong, J. (2014). GeoComputation for Geospatial Big Data. *Transactions in GIS*, 18, 1–2. Advance online publication. doi:10.1111/tgis.12131
- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large. *ACM SIGMOD International Conference on Management of Data*. 10.1145/233269.233324
- Zhang, X., Mei, C., Chen, D., & Li, J. (2016). Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognition*, 56, 1–15. Advance online publication. doi:10.1016/j.patcog.2016.02.013

Zhang, Y. Q., & Rajapakse, J. C. (2008). *Machine Learning in Bioinformatics*. Machine Learning in Bioinformatics. doi:10.1002/9780470397428

Zhao, W., Ma, H., & He, Q. (2009). Parallel K-means clustering based on MapReduce. Lecture Notes in Computer Science. doi:10.1007/978-3-642-10665-1_71

ADDITIONAL READING

Aggarwal, C. C., & Reddy, C. K. (2013). *DATA Clustering Algorithms and Applications*. Taylor & Francis Group. doi:10.1201/b15410

Chen, W. Y., Song, Y., Bai, H., Lin, C. J., & Chang, E. Y. (2011). Parallel spectral clustering in distributed systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3), 568–586. Advance online publication. doi:10.1109/TPAMI.2010.88 PMID:20421667

Freiknecht, J., Papp, S., Freiknecht, J., & Papp, S. (2018). Hadoop. Big Data in der Praxis. doi:10.3139/9783446456013.003

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264–323. doi:10.1145/331499.331504

KEY TERMS AND DEFINITIONS

BIRCH: This is a clustering algorithm which deals with large datasets.

DBSCAN: It's an algorithm based on grouping together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points.

GPU: GPU stands for graphics processing unit is a single chip processor mainly used for fast processing.

High Dimension: A big data is considered to be high dimensional as it contains many features.

Machine Learning: It is an application of artificial intelligence that incorporate the ability to learn automatically in the system.

Parallel Clustering: Fast processing can be done through parallel clustering and it can be done with the help of GPU.

PK-Means: It is an algorithm for gene clustering.

Challenges in Big Data Analysis

5

M. Govindarajan

Annamalai University, India

INTRODUCTION

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, Thomson Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges.

To handle the challenges, various computational complexities, information security, and computational method are needed to be known to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Big Data problems such as Heterogeneity, Scalability, Storage and Processing Issues, Noise Accumulation, Spurious Correlation, Incidental Endogeneity, Data Access and Sharing of Information, Quality of Data, Fault Tolerance, Speed/Velocity, Accuracy, Privacy and Security, Garbage Mining, Timeliness, Human Collaboration are to be addressed to design effective statistical procedures for exploring and predicting big data. These issues are discussed briefly in the following subsections.

BACKGROUND

David Lazer et al., (2009) discusses an emerging field that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors. Stadler et al., (2010) developed an efficient EM algorithm for numerical optimization with provable convergence properties. High dimensionality also gives rise to incidental endogeneity, a phenomenon that many unrelated covariates may incidentally be correlated with the residual noises. The endogeneity creates statistical biases and causes model selection inconsistency that lead to wrong scientific discoveries (Liao and Jiang, 2011; Fan and Liao, 2012). Jianqing Fan et al., (2013) gives overviews on the salient features of Big Data and how these features impact on paradigm change on statistical and computational methods as well as computing architectures. Nawsher Khan et al., (2014) comprehensively surveys and classifies the various attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security. This study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Future research directions in this field are determined based on opportunities and several open issues in Big Data domination. These research directions facilitate the exploration of the domain and the development of optimal techniques to address Big Data. Lenka Venkata Satyanarayana

DOI: 10.4018/978-1-7998-3479-3.ch041

(2015) provides an in-depth analysis of different platforms available for performing big data analytics. This paper surveys different hardware platforms available for big data analytics and assesses the advantages and drawbacks of Big Data. D. P. Acharjya et al., (2016) explore the potential impact of big data challenges, open research issues, and various tools associated with it. Akhil et al., (2017) analyzed the potential effect of big data challenges, open research issues, and different tools related with it. Ripon Patgiri (2018) presents a study report on numerous research issues and challenges of Big Data which is employed in very large dataset. Reihaneh H. Hariri et al., (2019) reviews previous work in big data analytics and presents a discussion of open challenges and future directions for recognizing and mitigating uncertainty in this domain.

FOCUS OF THE ARTICLE

The purpose of this chapter is to highlight the Big Data challenges and also provide a brief description of each challenge.

Heterogeneity

Big data are often created via aggregating many data sources corresponding to different sub-populations. Each sub-population might exhibit some unique features not shared by others. In classical settings where the sample size is small or moderate, data points from small sub-populations are generally categorized as “outliers” and it is hard to systematically model them due to insufficient observations. However, in the Big data era, the large sample size enables us to better understand heterogeneity, shedding lights toward studies such as exploring the association between certain covariates (e.g., genes or SNPs) and rare outcomes (e.g., rare diseases or diseases in small populations) and understanding why certain treatments (e.g. chemotherapy) benefit a subpopulation and harm another subpopulation. In short, the main advantage brought by Big Data is to understand heterogeneity of sub-populations such as the benefits of certain personalized treatments, which are infeasible when sample size is small or moderate.

Scalability

The processor technology has changed in recent years. The clock speeds have largely stalled and processors are being built with more number of cores instead. Previously data processing systems had to worry about parallelism across nodes in a cluster but now the concern has shifted to parallelism within a single node. In past the techniques which were used to do parallel data processing across data nodes aren't capable of handling intra-node parallelism. This is because of the fact that many more hardware resources such as cache and processor memory channels are shared across a core in a single node.

The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks. There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having

the same performance between sequential and random data transfer. Thus what kind of storage devices are to be used is again a big question for data storage.

Storage and Processing Issues

The storage available is not enough for storing the large amount of data which is being produced by almost everything: Social Media sites are themselves a great contributor along with the sensor devices etc. Because of the rigorous demands of the Big data on networks, storage and servers outsourcing the data to cloud may seem an option. Uploading this large amount of data in cloud doesn't solve the problem. Since Big data insights require getting all the data collected and then linking it in a way to extract important information. Terabytes of data will take large amount of time to get uploaded in cloud and moreover this data is changing so rapidly which will make this data hard to be uploaded in real time. At the same time, the cloud's distributed nature is also problematic for Big data analysis. Thus the cloud issues with Big Data can be categorized into Capacity and Performance issues.

The transportation of data from storage point to processing point can be avoided in two ways. One is to process in the storage place only and results can be transferred or transport only that data to computation which is important. But both these methods would require integrity and provenance of data to be maintained. Processing of such large amount of data also takes large amount of time. To find suitable elements whole of data Set needs to be Scanned which is somewhat not possible. Thus Building up indexes right in the beginning while collecting and storing the data is a good practice and reduces processing time considerably.

Noise Accumulation

Analyzing Big Data requires us to simultaneously estimate or test many parameters. These estimate errors accumulate when a decision or prediction rule depends on a large number of such parameters. Such a noise accumulation effect is especially severe in high dimensions and may even dominate the true signals. It is usually handled by the sparsity assumption (Donoho, 2000; Hastie et al., 2009; Buhlmann and van der Geer, 2011). Take high-dimensional classification for instance. Poor classification is due to the existence of many weak features that do not contribute to the reduction of classification error (Fan and Fan, 2008). In other words, variable selection plays a pivotal role in overcoming noise accumulation in classification and regression prediction. However, variable selection in high dimensions is challenging due to spurious correlation, incidental endogeneity, heterogeneity, and measurement errors.

Spurious Correlation

High dimensionality also brings spurious correlation, referring to the fact that many uncorrelated random variables may have high sample correlations in high dimensions. Spurious correlation may cause false scientific discoveries and wrong statistical inferences. Besides variable selection, spurious correlation may also lead to wrong statistical inference.

Incidental Endogeneity

Incidental endogeneity is another subtle issue raised by high dimensionality. Unlike spurious correlation, incidental endogeneity refers to the genuine existence of correlations between variables unintentionally,

both due to high-dimensionality. The former is analogous to find two persons look alike but have no genetic relation, whereas the latter is similar to bumping into an acquaintance, both easily occurring in a big city. More generally, endogeneity happens as a result of selection biases, measurement errors, and omitted variables.

Data Access and Sharing of Information

If data is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner. This makes the Data management and governance process bit complex adding the necessity to make Data open and make it available to government agencies in standardized manner with standardized APIs, metadata and formats thus leading to better decision making, business intelligence and productivity improvements. Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness.

Quality of Data

Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Business Leaders will always want more and more data storage whereas the IT Leaders will take all technical aspects in mind before storing all the data. Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

COMPLEXITY

Fault Tolerance

With the incoming of new technologies like Cloud computing and Big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault-tolerant computing is extremely hard, involving intricate algorithms. It is simply not possible to devise absolutely foolproof, 100% reliable fault tolerant machines or software. Thus the main task is to reduce the probability of failure to an “acceptable” level. Unfortunately, the more this probability is reduced, the higher the cost.

Two methods which seem to increase the fault tolerance in Big data are as: First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation. One node is assigned the work of observing that these nodes are working properly. If something happens that particular task is restarted.

But sometimes it's quite possible that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the input of the previous task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

Speed/Velocity

For big data, speed/velocity really matters. The capability of fast accessing and mining big data is not just a subjective desire, it is an obligation especially for data streams (a common format of big data) – a processing/mining task must be finished within a certain period of time, otherwise, the processing/mining results becomes less valuable or even worthless. Exemplary applications with real-time requests include earthquake prediction, stock market prediction and agent-based autonomous exchange (buying/selling) systems. Speed is also relevant to scalability – conquering or partially solving anyone helps the other one. The speed of data mining depends on two major factors: data access time (determined mainly by the underlying data system) and, of course, the efficiency of the mining algorithms themselves. Exploitation of advanced indexing schemes is the key to the speed issue.

An additional approach to boost the speed of big data access and mining is through maximally identifying and exploiting the potential parallelism in the access and mining algorithms. The elasticity and parallelism support of cloud computing are the most promising facilities for boosting the performance and scalability of big data mining systems. It is interesting to note that the MapReduce parallel computing model is applicable to only a rather limited class of data-intensive computing problems. Therefore, design of new and more efficient parallel computing models besides MapReduce is greatly desired, but calls for really creative minds.

Accuracy

Data analysis is typically buoyed by relatively accurate data obtained from structured databases with limited sources. Therefore, such analysis results are accurate. However, analysis is adversely affected by the increase in the amount of and the variety in data sources with data volume (Che et al., 2013). In data stream scenarios, high-speed data strongly constrain processing algorithms spatially and temporally. Hence, stream-specific requirements must be fulfilled to process these data (Bifet et al., 2010).

Privacy and Security

It is the most important issue with big data which is sensitive and includes conceptual, technical as well as legal significance.

- The personal information of a person when combined with external large data sets leads to the inference of new facts about that person and it's possible that these kinds of facts about the person are secretive and the person might not want the Data Owner to know or any person to know about them.
- Information regarding the users (people) is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.
- Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.
- Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

Garbage Mining

In the big data epoch, the amount of data produced and inhabited on the World Wide Web keeps increasing at an amazingly fast pace. In such an environment, data can quickly become obsolete, contaminated, and inadequate. In addition, there are data created as junks like junk emails. For having a relatively clean cyberspace and clean World Wide Web, attentions and research efforts are required. Cyberspace cleaning is not an easy task because of at least two foreseeable reasons: garbage is hidden, and there is an ownership issue. Garbage mining is a serious research topic, different but related to big data mining – for the sake the sustainability of our digital environment, “mining for garbage” (and cleaning it) is as important as “mining for knowledge” (the canonical sense of data mining). This is especially so in the new era of big data. Garbage definition remains one of the greatest challenges

Timeliness

With size comes the issue of speed. With increasing volume of data the time to analyze the data will also increase. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data faster.

Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modelling and analysis phase in the pipeline. In today’s complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration (Harshawardhan et al., 2014).

FUTURE RESEARCH DIRECTIONS

The amount of data collected from various applications all over the world across a wide variety of fields today is expected to double every two years. It has no utility unless these are analyzed to get useful information. This necessitates the development of techniques which can be used to facilitate big data analysis. The development of powerful computers is a boon to implement these techniques leading to automated systems. The transformation of data into knowledge is by no means an easy task for high performance large-scale data processing, including exploiting parallelism of current and upcoming computer architectures for data mining. Moreover, these data may involve uncertainty in many different forms. Many different models like fuzzy sets, rough sets, soft sets, neural networks, their generalizations and hybrid models obtained by combining two or more of these models have been found to be fruitful in representing data. These models are also very much fruitful for analysis. More often than not, big data are reduced to include only the important characteristics necessary from a particular study point of view or depending upon the application area. So, reduction techniques have been developed. Often

the data collected have missing values. These values need to be generated or the tuples having these missing values are eliminated from the data set before analysis. More importantly, these new challenges may comprise, sometimes even deteriorate, the performance, efficiency and scalability of the dedicated data intensive computing systems. The later approach sometimes leads to loss of information and hence not preferred. This brings up many research issues in the industry and research community in forms of capturing and accessing data effectively. In addition, fast processing while achieving high performance and high throughput, and storing it efficiently for future use is another issue. Further, programming for big data analysis is an important challenging issue. Expressing data access requirements of applications and designing programming language abstractions to exploit parallelism are an immediate need (Acharjya et al., 2015). Additionally, machine learning concepts and tools are gaining popularity among researchers to facilitate meaningful results from these concepts. Research in the area of machine learning for big data has focused on data processing, algorithm implementation, and optimization. Many of the machine learning tools for big data are started recently needs drastic change to adopt it. More efficient tools can be developed for dealing with problems inherent to big data depending upon the advantages and limitations of each tool. The efficient tools to be developed must have provision to handle noisy and imbalance data, uncertainty and inconsistency, and missing values.

CONCLUSION

This chapter provides an overview of challenges in big data analysis. The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts (Berkovich, S., Liao, D., 2012). Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights in the target domain. However, this brings a series of new challenges to the research community. From a security perspective, the major concerns of Big Data are privacy, integrity, availability, and confidentiality with respect to outsourced data. Large amounts of data are stored in cloud platforms. However, customers cannot physically check the outsourced data. Thus, data integrity is jeopardized. Given the lack of data support caused by remote access and the lack of information regarding internal storage, integrity assessment is difficult. Big Data involves large systems, profits, and challenges. Therefore, additional research is necessary to improve the efficiency of integrity evaluation online, as well as the display, analysis, and storage of Big Data.

REFERENCES

- Bhosale & Gadekar. (2014). A review paper on Big Data and Hadoop. *International Journal of Scientific and Research Publications*, 4(10), 1–7.
- Acharjya, D. P., Dehuri, S., & Sanyal, S. (2015). *Computational Intelligence for Big Data Analysis*. Springer International Publishing AG. doi:10.1007/978-3-319-16598-1
- Acharjya, D. P., & Kauser Ahmed, P. (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications*, 7(2), 511–518.

- Akhil, Shravya, & Uma. (2017). Survey on the Challenges And Issues On Big Data Analytics. *International Journal of Mechanical Engineering and Technology*, 8(12), 138–149.
- Berkovich, S., & Liao, D. (2012). On Clusterization of big data Streams. In *3rd International Conference on Computing for Geospatial Research and Applications*, ACM Press.
- Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). Moa:Massive online analysis. *Journal of Machine Learning Research*, 11, 1601–1604.
- Bühlmann, P., & van der Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. doi:10.1007/978-3-642-20192-9
- Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data Mining: challenges, issues, and opportunities. In *Database Systems for Advanced Applications* (pp. 1–15). Springer. doi:10.1007/978-3-642-40270-8_1
- Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–32.
- Donoho, D., & Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via L1 minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5), 2197–2202.
- Fan, Hany, & Liuz. (2013). *Challenges of Big Data Analysis*. arXiv:1308.1479v1 [stat.ML]
- Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6), 2605–2637. doi:10.1214/07-AOS504 PMID:19169416
- Fan, J., & Liao, Y. (2012). *Endogeneity in ultrahigh dimension*. Technical report.
- Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data*, 6(1), 44. doi:10.1186/40537-019-0206-3
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. doi:10.1007/978-0-387-84858-7
- Khan, Yaqoob, Hashem, & Inayat, Ali, Alam, Shiraz, & Gani. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal*, 1–18. PMID:25136682
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., & Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721–723. doi:10.1126/science.1167742 PMID:19197046
- Liao, Y., & Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *Annals of Statistics*, 39(6), 3003–3031. doi:10.1214/11-AOS930
- Patgiri, R. (2018). Issues and Challenges in Big Data: A Survey. *ICDCIT 2018. LNCS*, 10722, 295–300.
- Satyanarayana, L. V. (2015). A Survey on Challenges and Advantages in Big Data. *International Journal of Computer Science and Technology*, 6(2), 115–119.
- Stadler, N., Bühlmann, P., & Van De Geer, S. (2010). 1-penalization for mixture regression models. *Test*, 19(2), 209–256. doi:10.1007/11749-010-0197-z

KEY TERMS AND DEFINITIONS

Big Data: Big data hold great promises for discovering subtle population patterns and heterogeneities that are not possible with small-scale data.

Big Data Problems: Big data problems such as heterogeneity, noise accumulation, spurious correlations, and incidental endogeneity, in addition to balancing the statistical accuracy and computational efficiency.

Heterogeneity: Big data are often created via aggregating many data sources corresponding to different sub-populations.

Incidental Endogeneity: Incidental endogeneity is another subtle issue raised by high dimensionality.

Noise Accumulation: Analyzing Big Data requires us to simultaneously estimate or test many parameters. These estimate errors accumulate when a decision or prediction rule depends on a large number of such parameters.

Scalability: The scalability issue of big data has led towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters.

Spurious Correlation: High dimensionality also brings spurious correlation, referring to the fact that many uncorrelated random variables may have high sample correlations in high dimensions.

Big Data Analytics and IoT in Smart City Applications

Mamata Rath

 <https://orcid.org/0000-0002-2277-1012>

Birla School of Management, Birla Global University, India

INTRODUCTION

There are many Real Time Applications of Big Data Analytics in Smart City based systems, that incorporates Smart City idea, big data innovations, ongoing big data analytics, urban improvement, data and correspondence innovation, Internet of Things. Technological insurgency in the ongoing past has empowered the idea of Smart City for urban advancement. Smart City idea is imagined with the destinations of giving better administrations to the residents and improves the personal satisfaction. Data and Communication Technology (ICT) and Internet of Things (IoT) made smart city applications as a lot less difficult and compelling. Big data advances assume a significant job in smart city applications. This examination work gives an outline of the job of big data in structure smart city applications and proposes a system for continuous big data analytics. Constant big data analytics help in settling on better choices and progressively precise forecasts at opportune time to offer better administrations to the residents. Here, we focus on some of significant arrangements and administrations for the smart city where the continuous big data analytics and IoT helps in improving the nature of administrations in smart city applications.

There are learning models developed quality education in smart urban communities, that incorporates long transient memory systems, IoT smart city data examination, data and correspondence advancements, sustainable urban life, IoT based administrations, IoT data forecast, Internet of Things idea, profound learning strategies, profound learning methods, IoT based smart city applications, city partners, air quality expectation, profound learning model, smart city forecast problems. In ongoing years, Internet of Things (IoT) idea has turned into a promising exploration point in numerous zones including industry, trade and training. Smart urban areas utilize IoT based administrations and applications to make a sustainable urban life. By utilizing data and correspondence advances, IoT empowers smart urban communities to make city partners progressively mindful, intuitive and productive. With the expansion in number of IoT based smart city applications, the measure of data created by these applications is expanded enormously. Governments and city partners play it safe to process these data and anticipate future impacts to guarantee sustainable advancement. In expectation setting, profound learning procedures have been utilized for a few guaging issues in big data. This motivates us to utilize profound learning techniques for forecast of IoT data. Thus, in exploration work by many eminent researchers, novel profound learning models are proposed for breaking down IoT smart city data. The analysts present novel model dependent on Long Short Term Memory (LSTM) systems to foresee future estimations of air quality in a smart city. The assessment consequences of the proposed model are seen as promising and they demonstrate that the model can be utilized in other smart city forecast issues also.

DOI: 10.4018/978-1-7998-3479-3.ch042

Different Computing Platforms for Big Data Analytics in Electric Vehicle Infrastructures have been developed that incorporate data-concentrated investigation, carbon impression, SG mix, figuring stages, dispersed distributed computing, EV joining, smart vehicular applications, omnipresent arrangement, IoT gadgets, smart matrix mix, wise transportation framework, transport situated smart urban areas, Big Data analytics stages, EV rollout, BDA exercises, structural layers, electric vehicle foundations, ITS, TOSC, appropriated edge-haze computing. With the development of consistently developing smart vehicular applications and pervasive sending of IoT gadgets crosswise over various engineering layers of Intelligent Transportation System (ITS), data-escalated examination rises to be a significant test. Without incredible correspondence and computational help, different vehicular applications and administrations will at present remain in the idea stage and can't be tried in the daily life. In this examination work, the specialists think about the instance of Electric Vehicle (EV) to Smart Grid (SG) combination. The EVs are key players for Transport Oriented Smart Cities (TOSC) as they help urban areas to end up greener by lessening discharges and carbon impression. The analysts break down various use-cases in EV to SG combination to demonstrate how Big Data Analytics (BDA) stages can assume an essential job towards effective EV rollout. The analysts at that point present two figuring stages in particular, dispersed distributed computing and edge/haze registering. The scientists featured the distinctive highlights of each towards supporting BDA exercises in EV reconciliation. At long last, the specialists give a detailed outline of chances, patterns, and difficulties of both these processing strategies.

Big data analytics require advances and technical mechanism that can change a lot of organized, unstructured, and semi-organized data into a more reasonable data and metadata design for explanatory procedures. The calculations utilized as a part of these explanatory instruments must find examples, patterns, and relationships over an assortment of time skylines in the data. In the wake of breaking down the data, these instruments envision the discoveries in tables, diagrams, and spatial outlines for proficient decision making. Along these lines, big data investigation is a genuine test for some applications due to data unpredictability and the adaptability of basic calculations that help such procedures. acquiring supportive data from big data investigation is a basic issue that requires adaptable logical calculations and strategies to return all around planned outcomes, though current methods and calculations are wasteful to deal with big data analytics. In this manner, huge framework and extra applications are important to help data parallelism. In addition, data sources, for example, rapid data stream got from various data sources, have diverse configurations, which makes coordinating different hotspots for analytics arrangements basic. Subsequently, the assessment is centered around the execution of current calculations utilized as a part of big data investigation, which isn't rising directly with the fast increment in computational assets.

By and large, when the expression "big data," is coined, people quickly think about enormous data volumes. In case these were the main reason behind medicinal services administrations to get the better approach for taking care of data, they could regulate without, in light of the way that most of them could contain what they have in a solid social database. In 2001, Doug Laney depicted the "3 Vs" of enormous and complex data as "Volume, Velocity and Variety.". While restorative administrations CIOs could benefit by all of the three, the highlight should be on combination. This is an example experienced by human administrations, just as by all endeavors. Figure 1 demonstrates Big data applications in different fields. Beside customary patient data contained in substance, there are various pictures and sounds recorded, from x-shafts and ultrasounds, to Doppler and MRI imaging. A couple of experts particularly need that their talks with patients be recorded for the patient's bit of leeway. This get-together of dissimilar data is generally unstructured and can't be mentioned in the immaculate tables and sections of a social database. This is the spot big databases, as Hadoop, score. In any case, it is one thing to store big data and extremely another to recoup it truly. Data analysts who can design procedures to remove significant data from the non-successive and evidently unpredictable big databases are presently popular.

Figure 1. Applications of Big data Analytics in various fields

Security and Banking	Manufacturing and Production
Communications and Media	Education & Training
Health Care Applications	Energy Sectors and Utilities
Transportations	Home Appliances

These techniques are challenging and hard to solve, however the IT business is beginning to convey activities that make significant data extraction less demanding. There is additionally a move to a half breed database structure, where data is put away in both a social and a “NoSQL” database. Where social insurance elements have handled this obstacle, the outcome is an all encompassing perspective of the patient, which expels a portion of the unpredictability of finding for the medicinal professional and makes life less difficult for the patient. It likewise opens the path for the move to machine-to-machine (M2M) correspondence and the utilization of computerized reasoning to filter through and investigate data transmitted from the sensors gathering it. The future guarantee is examination that will screen well-being more than ever however there are likewise further issues that should be tended to, for example, data protection and security. Figure 2 represents various smart applications in smart city using Internet of Things.

Figure 2. Applications of IoT in Smart City

Applications of IoT in Smart City	
Smart Traffic Management	Smart Industry
Smart Energy and Smart Grid	Smart healthcare
Smart Transportations	Smart Home Appliances

For analysis of major type of business process, Big Data helps in transforming major selling practice by improved techniques and accurate analysis of presented data. As shown in fig.1 some of such process are Banking and Security, Media and communication, Health Care Sector, Manufacturing, Training and Education, Insurance, Transportation, Energy and utilities. Using big data analytics in all these fields improves the quality of the process and results improved output. The Organisation of the paper has been done as follows. The Second section describes the Literature Review. Section 3 presents applications of big data analytics and IoT in smart city and at last Section 4 concludes the chapter.

2. BACKGROUND

This section presents background of the topic with literature study. Manjunatha, B. Annappa (2018) present Real Time Big Data Analytics in Smart City Applications, that incorporates Smart City idea, big data innovations, ongoing big data analytics, urban improvement, data and correspondence innovation, Internet of Things. Technological insurgency in the ongoing past has empowered the idea of Smart City for urban advancement. Smart City idea is imagined with the destinations of giving better administrations to the residents and improves the personal satisfaction. Data and Communication Technology (ICT) and Internet of Things (IoT) made smart city applications as a lot less difficult and compelling. Big data advances assume a significant job in smart city applications. This examination work gives an outline of the job of big data in structure smart city applications and proposes a system for continuous big data analytics. Constant big data analytics help in settling on better choices and progressively precise forecasts at opportune time to offer better administrations to the residents. Here, the analysts talked about some of significant arrangements and administrations for the smart city where the continuous big data analytics helps in improving the nature of administrations in smart city applications.

M. U. Azimayek (2017) present A profound learning model for air quality expectation in smart urban communities, that incorporates LSTM, long transient memory systems, IoT smart city data examination, data and correspondence advancements, sustainable urban life, IoT based administrations, IoT data forecast, Internet of Things idea, profound learning strategies, profound learning methods, IoT based smart city applications, city partners, air quality expectation, profound learning model, smart city forecast problems. In ongoing years, Internet of Things (IoT) idea has turned into a promising exploration point in numerous zones including industry, trade and training. Smart urban areas utilize IoT based administrations and applications to make a sustainable urban life. By utilizing data and correspondence advances, IoT empowers smart urban communities to make city partners progressively mindful, intuitive and productive. With the expansion in number of IoT based smart city applications, the measure of data created by these applications is expanded enormously. Governments and city partners play it safe to process these data and anticipate future impacts to guarantee sustainable advancement. In expectation setting, profound learning procedures have been utilized for a few gauging issues in big data. This motivates us to utilize profound learning techniques for forecast of IoT data. Thus, in this exploration work, a novel profound learning model is proposed for breaking down IoT smart city data. The analysts propose a novel model dependent on Long Short Term Memory (LSTM) systems to foresee future estimations of air quality in a smart city. The assessment consequences of the proposed model are seen as promising and they demonstrate that the model can be utilized in other smart city forecast issues also. Table 1. Illustrates details of literature survey carried out in this research work.

B. Cheng, S. Longo, F. Cirillo, M. Bauer, E. Kovacs (2015) present Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander, that incorporates Big Data stage, Internet of Things, IoT, human exercises, setting mindfulness, framework design, live city data and analytics stage, CiDAP, SmartSantander, smart city data platforms. The Internet of Things (IoT) is currently forming exploration work 's urban areas to make them increasingly associated, helpful, and keen. In any case, this change will profoundly depend on extricated qualities and bits of knowledge from the big data created by research work 's urban areas by means of sensors, gadgets, and human exercises. Many existing examinations and tasks have been done to make research work 's urban communities smart, concentrating more on the most proficient method to send different sensors and gadgets and after that gather data from them. Be that as it may, this is only the initial move towards smart urban areas and following stage will be to utilize the gathered data and empower setting mindfulness and insight into a wide range of uses

Table 1. Details of literature survey carried out in this research work

Literature Study Reference	year	Research Topic	Associated Concepts related to Respective Research
Manjunatha, B. Annappa	2018	Real Time Big Data Analytics in Smart City Applications	Smart City concept, big data technologies, real-time big data analytics, urban development, information and communication technology, Internet of Things
M. U. Azimayek	2017	A deep learning model for air quality prediction in smart cities	LSTM, long short term memory networks, IoT smart city data analysis, information and communication technologies, sustainable urban life, IoT based services, IoT data prediction, Internet of Things concept, deep learning methods, deep learning techniques, IoT based smart city applications, city stakeholders, air quality prediction, deep learning model, smart city prediction problems
B. Cheng, S. Longo, F. Cirillo, M. Bauer, E. Kovacs	2015	Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander	Big Data platform, Internet of Things, IoT, human activities, context-awareness, system architecture, live city data and analytics platform, CiDAP, SmartSantander, smart city data platforms
M. M. Hussain, M. M. S. Beg, M. S. Alam, M. Krishnamurthy, Q. M. Ali	2018	Computing Platforms for Big Data Analytics in Electric Vehicle Infrastructures	data-intensive analysis, carbon footprint, SG integration, computing platforms, distributed cloud computing, EV integration, smart vehicular applications, ubiquitous deployment, IoT devices, smart grid integration, intelligent transportation system, transport oriented smart cities, Big Data analytics platforms, EV rollout, BDA activities, architectural layers, electric vehicle infrastructures, ITS, TOSC, distributed edge-fog computing
S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. Ben Yahia, D. Draheim	2019	Towards Disaster Resilient Smart Cities: Can Internet of Things and Big Data Analytics Be the Game Changers?	IoT, big data analytics technologies, BDA technologies, disaster management activities, smart city incentives, data harvesting, data aggregation, data pre-processing, smart buildings, city pollution, natural disasters, disaster resilient smart cities, Internet of Things, DRSC environment, Hadoop Ecosystem, Spark, traffic simulator, Twitter, emergency evacuation path
M. R. Bashir, A. Q. Gill	2016	Towards an IoT Big Data Analytics Framework: Smart Buildings Systems	IoT Big Data analytics, IBDA framework, smart building system, data storage, data analysis
P. Bellini, P. Nesi, M. Paolucci, I. Zaza	2018	Smart City Architecture for Data Ingestion and Analytics: Processes and Solutions	provide support, decision support systems, interoperability, data ingestion, Sii-Mobility national smart city project, Km4City ontology, smart city data aggregation, service production, smart City architecture, smart city architectures, data analytics processes, Internet of everything, sensors, IOT
A. Sharif, J. Li, M. Khalil, R. Kumar, M. I. Sharif, A. Sharif	2017	IoT based smart traffic management system for smart cities using big data analytics	STS, smart city, smart traffic management system, Internet of Things, IoT, Smart Traffic System, smart cities, predictive analytics, traffic density, analytical scriptures, big data analytics, data processing, public traffic data, low cost vehicle, traffic update, public traffic management
B. Ahlgren, M. Hidell, E. C. -. Ngai	2016	Internet of Things for Smart Cities: Interoperability and Open Data	Internet of Things, smart cities, interoperability, Big Data analytics, open data accessibility, cloud services, GreenIoT platform, Uppsala, Sweden
J. Chin, V. Callaghan, I. Lam	2017	Understanding and personalising smart city services using machine learning, The Internet-of-Things and Big Data	personalising smart city services, machine learning, Internet-of-Things, Big Data, artificial intelligence, IoT, ML classification algorithms, Bayes Network, Naïve Bayesian, NB, nearest neighbour, NN, weather data, UK MetOffice, Transport for London, decision tree J48 algorithm

continued on following page

Table 1. Continued

Literature Study Reference	year	Research Topic	Associated Concepts related to Respective Research
J. Zenkert, M. Dornhofer, C. Weber, C. Ngoukam, M. Fathi	2018	Big data analytics in smart mobility: Modeling and analysis of the Aarhus smart city dataset	intelligent services, smart urban services, intelligent networking, sensing technologies, Open Data Aarhus datasets, sensor data, city Aarhus, smart mobility, Aarhus smart city dataset, information and communication technologies, Big Data analytics, Internet of Things, IoT, R Shiny application, road traffic data analysis, pollution data analysis, MapReduce framework, Hadoop cluster
P. Manjunath, M. Prakruthi, P. Gajkumar Shah	2018	IoT Driven with Big Data Analytics and Block Chain Application Scenarios	Internet of Things, big data analytics, block chain application scenarios
S. Shukla, Balachandran K, Sumitha V S	2016	A framework for smart transportation using Big Data	Big Data, information technology, data driven decision, India, smart city project, smart transportation systems, smart traffic signals
P. Rizwan, K. Suresh, M. R. Babu	2016	Real-time smart traffic management system for smart cities by using Internet of Things and big data	real-time smart traffic management system, smart cities, Internet of Things, Big Data, STMS, traffic indicators, traffic details, low-cost vehicle detecting sensors, IoT, traffic data, real time streaming data, traffic density analysis, predictive analytics, mobile application, user interface
Y. Sun, H. Song, A. J. Jara, R. Bie	2016	Internet of Things and Big Data Analytics for Smart and Connected Communities	Italy, Trento, cultural heritage, smart tourism, TreSight, mobile crowdsensing, cyber-physical cloud computing, smart sensors, ubiquitous network, IoT, SCC, big data analytics, Internet of Things
V. Dattana, K. Gupta, A. Kush	2019	A Probability based Model for Big Data Security in Smart City	big data security, smart technologies, economic data, smart city data, smart applications, smart city critical data, data objects, transport data, data analytics, smart traffic management system, sensitive data sharing, data leakage detection, highly sensitive data, probability based model, personal data, organisational data, environment data, energy data, crisis response, emergency management, disaster resilience, bigraph, critical data security
S. N. Shukla, T. A. Champaneria	2017	Survey of various data collection ways for smart transportation domain of smart city	smart city, IoT, M2M interaction, Information and Communication Technology, quality of life, smart transportation, smart transportation domain, data collection ways
J. Lieberman, A. Leidner, G. Percivall, C. RÄ¶nsdorf	2017	Using big data analytics and IoT principles to keep an eye on underground infrastructure	Concept Development Study, Open Geospatial Consortium, high-quality feature data, underground urban infrastructure, nonvisual methods, UGI geodata, hidden features, diverse velocity sensing streams, high-velocity sensing streams, realistic predictive models, lower construction costs, efficient infrastructure operation, sound disaster preparedness, smart city services, IoT principles, Sensor Web observation standards, functional digital twins, hidden infrastructure, underground built environment, eye, underground infrastructure, Big Data analytics, OGC geodata
S. Nguyen, Z. Salcic, X. Zhang	2018	Big Data Processing in Fog - Smart Parking Case Study	Hadoop MapReduce, fog - smart parking, fog computing, Big Data processing, sensor-based system, data analytics tasks, data analytics system, monitoring carpark occupancy, IoT-based platform, vehicle parking, traffic congestion, traffic efficiency
C. Costa, M. Y. Santos	2016	BASIS: A big data architecture for smart cities	Big data architecture, smart cities, Internet of Things, IoT, BASIS, multiple abstraction layers, public data availability

continued on following page

Table 1. Continued

Literature Study Reference	year	Research Topic	Associated Concepts related to Respective Research
S. V. Nandury, B. A. Begum	2017	Big data for smart grid operation in smart cities	smart grid operation, conventional power grids, smart city environment, smart grid applications, Big data analytics, smart grid management, smart grid maintenance, power generation, power consumption, power distribution, SWIFT, ICT backbone architecture
P. Nesi, G. Pantaleo, M. Paolucci, I. Zaza	2018	Auditing and Assessment of Data Traffic Flows in an IoT Architecture	data traffic flows, Internet of Things, Smart City environments, flexible smart living lab, Smart City data, automated decision-making processes, data flows, connected IoT devices, data-driven applications, data shadow, IoT Smart City Architecture, Snap4City framework, DevDash tools, AMMA tools
M. E. ARASS, I. Tikito, N. Souissi	2018	An Audit Framework for Data Lifecycles in a Big Data context	Internet of Things, Big data lifecycles, data usage context, data collection, data management, Big data context, audit framework
M. M. Rathore, A. Ahmad, A. Paul	2016	IoT-based smart city development using big data analytical approach	IoT-based smart city development, Big Data analytical approach, urban public development, city development, Internet, smart city system, smart home sensors, vehicular networking, weather sensors, water sensors, smart parking sensors, surveillance objects, Hadoop ecosystem, data generation, data collection, data aggregation, data classification, data filtration, data preprocessing, spark
P. J. Navarathna, V. P. Malagi	2018	Artificial Intelligence in Smart City Analysis	AI, fully functional smart city, artificial intelligence, sustainable development, growing fields, conventional cities, smart cities

and administrations by means of an adaptable big data stage. In this examination work, the scientists present the framework engineering and the significant structure issues of a live City Data and Analytics Platform, in particular CiDAP. All the more critically, the scientists offer research work ‘ s experience and exercises gained from structure this reasonable framework for an enormous scale running smart city proving ground, SmartSantander. Research work ‘ s work gives a significant guide to future Smart City stage planners with the goal that they can anticipate some training issues and allude to research work ‘ s arrangement when building their very own smart city data stages.

M. M. Hussain, M. M. S. Ask, M. S. Alam, M. Krishnamurthy, Q. M. Ali (2018) present Computing Platforms for Big Data Analytics in Electric Vehicle Infrastructures, that incorporates data-concentrated investigation, carbon impression, SG mix, figuring stages, dispersed distributed computing, EV joining, smart vehicular applications, omnipresent arrangement, IoT gadgets, smart matrix mix, wise transportation framework, transport situated smart urban areas, Big Data analytics stages, EV rollout, BDA exercises, structural layers, electric vehicle foundations, ITS, TOSC, appropriated edge-haze computing. With the development of consistently developing smart vehicular applications and pervasive sending of IoT gadgets crosswise over various engineering layers of Intelligent Transportation System (ITS), data-escalated examination rises to be a significant test. Without incredible correspondence and computational help, different vehicular applications and administrations will at present remain in the idea stage and can’t be tried in the daily life. In this examination work, the specialists think about the instance of Electric Vehicle (EV) to Smart Grid (SG) combination. The EVs are key players for Transport Oriented Smart Cities (TOSC) as they help urban areas to end up greener by lessening discharges and carbon impression. The analysts break down various use-cases in EV to SG combination to demonstrate how Big Data Analytics (BDA) stages can assume an essential job towards effective EV rollout. The analysts at that

point present two figuring stages in particular, dispersed distributed computing and edge/haze registering. The scientists featured the distinctive highlights of each towards supporting BDA exercises in EV reconciliation. At long last, the specialists give a detailed outline of chances, patterns, and difficulties of both these processing strategies.

S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. Ben Yahia, D. Draheim (2019) present *Towards Disaster Resilient Smart Cities: Can Internet of Things and Big Data Analytics Be the Game Changers?*, that incorporates IoT, big data analytics advancements, BDA innovations, catastrophe the board exercises, smart city motivations, data reaping, data collection, data pre-handling, smart structures, city contamination, cataclysmic events, fiasco strong smart urban areas, Internet of Things, DRSC condition, Hadoop Ecosystem, Spark, traffic test system, Twitter, crisis clearing path. Disasters (characteristic or man-made) can be deadly to human life, nature, and framework. The ongoing headways in the Internet of Things (IoT) and the advancement in big data analytics (BDA) advances have given an open chance to grow exceptionally required debacle versatile smart city conditions. In this exploration work, the specialists propose and talk about the novel reference design and theory of a debacle versatile smart city (DRSC) through the mix of the IoT and BDA advances. The proposed engineering offers a nonexclusive answer for debacle the board exercises in smart city motivations. A mix of the Hadoop Ecosystem and Spark are checked on to build up a productive DRSC condition that supports both ongoing and disconnected investigation. The execution model of nature comprises of data collecting, data total, data pre-handling, and big data analytics and administration stage. An assortment of datasets (i.e., smart structures, city contamination, traffic test system, and twitter) are used for the approval and assessment of the framework to recognize and create alarms for a fire in a structure, contamination level in the city, crisis departure way, and the gathering of data about catastrophic events (i.e., seismic tremors and tidal waves). The assessment of the framework effectiveness is estimated as far as handling time and throughput that shows the exhibition predominance of the proposed engineering. In addition, the key difficulties confronted are distinguished and quickly talked about.

M. R. Bashir, A. Q. Gill (2016) present *Towards an IoT Big Data Analytics Framework: Smart Buildings Systems*, that incorporates IoT Big Data analytics, IBDA structure, smart structure framework, data stockpiling, data analysis. There is a developing enthusiasm for IoT-empowered smart structures. In any case, the capacity and examination of enormous measure of fast continuous smart structure data is a difficult undertaking. There are various contemporary Big Data the board innovations and progressed analytics procedures that can be utilized to manage this test. There is a requirement for a coordinated IoT Big Data Analytics (IBDA) system to fill the examination hole in the Big Data Analytics domain. This exploration work presents one such IBDA structure for the capacity and investigation of constant data produced from IoT sensors conveyed inside the smart structure. The underlying form of the IBDA structure has been created by utilizing Python and the Big Data Cloudera stage. The materialness of the structure is exhibited with the assistance of a situation including the investigation of continuous smart structure data for naturally dealing with the oxygen level, iridescence and smoke/dangerous gases in various pieces of the smart structure. The underlying outcomes demonstrate that the proposed structure is fit for the reason and appears to be valuable for IoT-empowered Big Data Analytics for smart city. Table 1 displays details of literature survey carried out in this research work.

The present expansion and development in the field of Internet of Things (IoT) are giving surprising potential in the course of the novel era of healthcare. The visualization of the healthcare is expansively supported, as it progresses the perfection of life and health of people, including a few health controls. The perpetual increment of the multifaceted IoT devices in health is comprehensively tried by difficulties, for example, controlling the IoT terminal hubs utilized for health observing, continuous information

handling and shrewd choice and occasion administration. Healthcare plan has been proposed by few researchers which are based with examination of vitality collecting for health observing sensors and the acknowledgment of Big Data investigation in healthcare.

In the choice of the most modern decade, there has been an expanding passion for huge information look into, particularly for health administrations applications. The requisition of the distributed computing and the Internet of Things (IoT) worldview in the healthcare field can convey a few chances to therapeutic IT, and specialists trust that it can essentially enhance healthcare administrations and add to its ceaseless and orderly advancement in a major information condition, for example, Industry 4.0 applications. In any case, the expected assets to oversee such information in a cloud-IoT condition are as yet a major test. In like manner, another model is also planned to streamline virtual machines choice (VMs) in cloud-IoT health administrations applications to productively deal with a major measure of information in incorporated industry 4.0. Industry 4.0 applications require to process and examine huge information, which originate from various sources, for example, sensor information, without human mediation. The proposed demonstrate plans to improve the execution of the healthcare frameworks by lessening the partners' demand execution time, streamlining the required stockpiling of patients' huge information and giving an ongoing information recovery component for those applications.

The inclination of universal frameworks is supported by the improvement and dynamic selection of the Internet of Things (IoT) devices and their empowering advancements. IoT has been appeared to have critical potential in high-chance Environment, Health, and Safety (EHS) enterprises. In these businesses, human lives are in question and IoT-based applications are prepared to offer protected, dependable, and effective arrangements because of their capacity to work at a fine granular level and give rich low-level data. Another research audits existing distributed research on IoT-based applications in high-hazard EHS businesses with particular prominence on healthcare industry, nourishment inventory network (FSC), mining and vitality enterprises (oil and gas and atomic), insightful transportation (e.g., associated vehicles), and building and foundation administration for crisis reaction tasks until 2016. Web of Things (WoT) offers a consistent stage to interface individuals and items to each other for improving and making our lives simpler. This vision conveys us from form based unified plans to a more circulated condition offering a tremendous measure of utilizations, for example, intelligent wearable devices, smart home, perceptive versatility, and savvy urban areas. Few research talks regarding immaterialness of IoT in healthcare and prescription by displaying an all encompassing design of IoT eHealth biological community. Healthcare is ending up progressively hard to oversee because of inadequate and less compelling healthcare administrations to meet the expanding requests of rising maturing populace with constant ailments. We suggest this requires a progress from the center driven treatment to understanding driven healthcare where every operator, for example, clinic, patient, and administrations are flawlessly associated with each other. This patient-driven IoT eHealth biological community needs a multi-layer design: (1) device, (2) mist figuring and (3) cloud to engage treatment of complex information regarding its assortment, speed, and inertness. This mist driven IoT design is trailed by different case cases of administrations and applications that are actualized on those layers. Those illustrations run from versatile health, helped living, e-drug, inserts, early cautioning frameworks, to populace checking in keen urban communities.

Over the most recent couple of years, the m-healthcare applications in light of Internet of Things (IoT) have given multi-dimensional highlights and ongoing administrations. These applications give a stage to a huge number of individuals to get health refreshes routinely for a healthier way of life. Enlistment of IoT devices in the healthcare condition have renewed numerous highlights of these applications. The huge information created by IoT devices in healthcare space is broke down on the cloud rather than exclusively depending on restricted capacity and calculation assets of handheld devices.

In respect to this unique situation, a cloud-driven IoT based M-healthcare checking illness diagnosing system is proposed which predicts the potential ailment with its level of seriousness. Key wordings are characterized to produce client situated health estimations by investigating the idea of computational sciences. The structural model for brilliant understudy healthcare is intended for application situation. The outcomes are registered subsequent to preparing the health estimations in a particular setting. In few research, deliberate understudy point of view health information is created utilizing UCI dataset and medicinal sensors to anticipate the understudy with various illness seriousness. Finding plans are connected utilizing different best in class order calculations and the outcomes are registered in light of precision, affectability, specificity, and F-measure. Exploratory outcomes demonstrate that the proposed technique beats the benchmark strategies for illness forecast.

3. BIG DATA ANALYTICS AND IOT IN SMART CITY APPLICATIONS

One of the biggest obstacles obstructing to utilize big data in social insurance is the manner by which restorative data is spread crosswise over numerous sources represented by various states, clinics, and authoritative offices. Combination of these data sources would require building up another foundation where all data suppliers team up with each other. Similarly critical is actualizing new data examination instruments and procedures. Social insurance needs to get up to speed with different ventures that have effectively moved from standard relapse based techniques to more future-arranged like prescient investigation, machine learning, and chart examination. Table 2 displays details of major contributions performed on applications of big data analytics and IoT in smart city applications.

Big data analytics considers the progression of a wide variety of uses. Things can be looked for, directed, analyzed, and even consolidated into aggregate beguilements. Ventures, human services, and urban regions are abusing IoT data driven frameworks to make these affiliations increasingly beneficial, along these lines, upgrading the lives of occupants. For making IoT a reality, data made by sensors, propelled cell phones, watches, and various wearables ought to be facilitated; furthermore, the significance of IoT data should be explicitly addressed. In any case, the Big Data nature of IoT data powers difficulties that ought to be tended to keeping in consideration the ultimate objective to give flexible and profitable IoT data driven establishments. These issues have been dealt with and focussed on the issues of depicting the centrality of IoT spilling data using reasoning and fusing this data in a learning outline.

The Internet-of-Things (IoT) has supposed control over the business range, and its applications fluctuate broadly from farming and social insurance to transportation. A hospital condition can be extremely upsetting, particularly for senior people and youngsters. With the consistently expanding total populace, the regular patient-physical check up has lost its viability. Thus, keen medicinal services turns out to be critical. Savvy social insurance can be actualized at all levels, beginning from temperature checking for children to following essential signs in the elderly. Data technology has progressed amid the most recent five decades to the phase where its effect is being felt by the general public in each administration that it gets from media, business, health care, shopper devices, vitality and power, and transportation domains. Amid this course of human-technology collaboration tremendous measure of data and learning exchange happens straightforwardly between specialist co-ops and their customers, and in addition by implication between customers. Since human propensity is to “investigate” its past with a specific end goal to foresee the “future”, monitoring this powerfully gushing voluminous heterogeneous data, called Big Data (BD), and examining it for significant revelation of learning that prompts esteem included business turns into a critical research action. It is in this setting research in Big Data (BD) figuring has

Table 2. Contributions of big data analytics and IoT in smart city

· Real Time Big Data Analytics in Smart City Applications
· A deep learning model for air quality prediction in smart cities
· Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander
· Computing Platforms for Big Data Analytics in Electric Vehicle Infrastructures
· Towards Disaster Resilient Smart Cities with Focus on Internet of Things and Big Data Analytics Be the Game Changer
· Towards an IoT Big Data Analytics Framework: Smart Buildings Systems
· Smart City Architecture for Data Ingestion and Analytics: Processes and Solutions
· IoT based smart traffic management system for smart cities using big data analytics
· Internet of Things for Smart Cities and Interoperability and Open Data
· Understanding and personalising smart city services using machine learning, The Internet-of-Things and Big Data
· Big data analytics in smart mobility: Modeling and analysis of the Aarhus smart city dataset
· IoT Driven with Big Data Analytics and Block Chain Application Scenarios
· A framework for smart transportation using Big Data
· Real-time smart traffic management system for smart cities by using Internet of Things and big data
· Internet of Things and Big Data Analytics for Smart and Connected Communities
· A Probability based Model for Big Data Security in Smart City
· Survey of various data collection ways for smart transportation domain of smart city
· Using big data analytics and IoT principles to keep an eye on underground infrastructure
· Big Data Processing in Fog - Smart Parking Case Study
· BASIS: A big data architecture for smart cities
· Big data for smart grid operation in smart cities
· Auditing and Assessment of Data Traffic Flows in an IoT Architecture
· An Audit Framework for Data Lifecycles in a Big Data context
· IoT-based smart city development using big data analytical approach
· Artificial Intelligence in Smart City Analysis

developed. Important choices can be construct just with respect to huge information disclosure, which thus requires a decent comprehension of the attributes of the aggregated data, a fitting grouping of this enormous gathering, and a proficient investigation of it. Health care part is a basic framework since its administrations influence the lives of humans and the absence of administration coherence might be unfortunate to the economy and human lives. The huge measure of data gathered by this part from its customers is organized into sections based on importance and relevance and further processed through intelligent utilities.

4. FUTURE RESEARCH DIRECTIONS

Big data analytics and IoT are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of

technologies and analysis will result in predicting the future drift of events. Main focus of this section is to discuss open research issues in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing. However it is not limited to these issues.

Big Data and IoT in Mobile Communication

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile devices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. In a broader sense, just like the internet, Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Conversely, it is still mystifying to understand IoT well, including definitions, content and differences from other similar concepts. Several diversified technologies such as computational intelligence, and big-data can be incorporated together to improve the data management and knowledge discovery of large scale automation applications. Much research in this direction has been carried out by eminent researchers.

Big Data and IoT in Cloud Computing

The development of virtualization technologies have made super computing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a salable and on demand availability of resources and data. Cloud computing harmonize massive data by on- demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big data and cloud computing are discussed in detail by many researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools.

Evolutionary Computing and Big Data Analytics

Bio-inspired computing is a technique inspired by nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism

search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications.

Big Data Analysis in IoT and Quantum Computing Research

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously [33]. This exponential improvement in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building quantum computer could soon be possible. Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave in quantum way. For example, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale quantum computers compared with classical computers. Hence it is a challenge for this generation to build a quantum computer and facilitate quantum computing to solve big data problems.

5. CONCLUSION

Big data analytics empowers data diggers and researchers to examine an extensive volume of data that may not be outfit utilizing customary apparatuses. Big data analytics require advances and statistical instruments that can change a lot of organized, unstructured, and semi-organized data into a more reasonable data and metadata design for explanatory procedures. There is tremendous positive potential concerning application of big data in smart city applications. The above chapter focuses on many versatile applications related to smart city and the contribution of big data analytics in IoT based systems to enhance the functionality of intelligent IoT systems.

REFERENCES

- Ahlgren, B., Hidell, M., & Ngai, E. C. (2016). Internet of Things for Smart Cities: Interoperability and Open Data. *IEEE Internet Computing*, 20(6), 52–56. doi:10.1109/MIC.2016.124
- Arass, M. E., Tikito, I., & Souissi, N.(2018). An Audit Framework for Data Lifecycles in a Big Data context. *International Conference on Selected Topics in Mobile and Wireless Networking (MoWNet)*, 1-5. doi: 10.1109/MoWNet.2018.8428883

- Bashir, M. R., & Gill, A. Q. (2016). Towards an IoT Big Data Analytics Framework: Smart Buildings Systems. *IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 1325-1332. 10.1109/HPCC-SmartCity-DSS.2016.0188
- Bellini, P., Nesi, P., Paolucci, M., & Zaza, I. (2018). Smart City Architecture for Data Ingestion and Analytics: Processes and Solutions. *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, 137-144. 10.1109/BigDataService.2018.00028
- Cheng, B., Longo, S., Cirillo, F., Bauer, M., & Kovacs, E. (2015). Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander. *IEEE International Congress on Big Data*, 592-599. 10.1109/BigDataCongress.2015.91
- Chin, J., Callaghan, V., & Lam, I. (2017). Understanding and personalising smart city services using machine learning, The Internet-of-Things and Big Data. *IEEE 26th International Symposium on Industrial Electronics (ISIE)*, 2050-2055. 10.1109/ISIE.2017.8001570
- Costa, C., & Santos, M. Y. (2016). BASIS: A big data architecture for smart cities. *SAI Computing Conference (SAI)*, 1247-1256. 10.1109/SAI.2016.7556139
- Dattana, V., Gupta, K., & Kush, A. (2019). A Probability based Model for Big Data Security in Smart City. *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, 1-6. 10.1109/ICBDSC.2019.8645607
- Hussain, M. M., Beg, M. M. S., Alam, M. S., Krishnamurthy, M., & Ali, Q. M. (2018). Computing Platforms for Big Data Analytics in Electric Vehicle Infrastructures. *4th International Conference on Big Data Computing and Communications (BIGCOM)*, 138-143. 10.1109/BIGCOM.2018.00029
- Kök, I., Şimşek, M. U., & Özdemir, S. (2017). A deep learning model for air quality prediction in smart cities. *2017 IEEE International Conference on Big Data (Big Data)*, 1983-1990. 10.1109/BigData.2017.8258144
- Lieberman, J., Leidner, A., Percivall, G., & Rösndorf, C. (2017). Using big data analytics and IoT principles to keep an eye on underground infrastructure. *IEEE International Conference on Big Data (Big Data)*, 4592-4601. 10.1109/BigData.2017.8258503
- Manjunath, P., Prakruthi, M., & Gajkumar Shah, P. (2018). IoT Driven with Big Data Analytics and Block Chain Application Scenarios. *Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, 569-572. 10.1109/ICGCIoT.2018.8752973
- Manjunatha & Annappa B. (2018). Real Time Big Data Analytics in Smart City Applications. *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 279-284.doi: 10.1109/IC3IoT.2018.8668106
- Nandury, S. V., & Begum, B. A. (2017). Big data for smart grid operation in smart cities. *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 1507-1511. 10.1109/WiSPNET.2017.8300013
- Navarathna, P. J., & Malagi, V. P. (2018). Artificial Intelligence in Smart City Analysis. *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 44-47. 10.1109/ICSSIT.2018.8748476

Nesi, P., Pantaleo, G., Paolucci, M., & Zaza, I. (2018). Auditing and Assessment of Data Traffic Flows in an IoT Architecture. *IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, 388-391. 10.1109/CIC.2018.00058

Nguyen, S., Salcic, Z., & Zhang, X. (2018). Big Data Processing in Fog - Smart Parking Case Study. *IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 127-134. doi:10.1109/BDCloud.2018.00031

Rathore, M. M., Ahmad, A., & Paul, A. (2016). IoT-based smart city development using big data analytical approach. *IEEE International Conference on Automatica (ICA-ACCA)*, 1-8. 10.1109/ICA-ACCA.2016.7778510

Rizwan, P., Suresh, K., & Babu, M. R. (2016). Real-time smart traffic management system for smart cities by using Internet of Things and big data. *International Conference on Emerging Technological Trends (ICETT)*, 1-7. . 2016.787366010.1109/ICETT.2016.7873660

Shah, S. A., Seker, D. Z., Rathore, M. M., Hameed, S., & Yahia, S. (2019). Towards Disaster Resilient Smart Cities: Can Internet of Things and Big Data Analytics Be the Game Changers? *IEEE Access*, 7, 91885-91903. doi:10.1109/ACCESS.2019.2928233

Sharif, A., Li, J., Khalil, M., Kumar, R., Sharif, M. I., & Sharif, A. (2017). Internet of things — smart traffic management system for smart cities using big data analytics. *2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 281-284. doi: 10.1109/ICCWAMTIP.2017.8301496

Shukla, S., Balachandran, K., & Sumitha, V. S. (2016). A framework for smart transportation using Big Data. *International Conference on ICT in Business Industry & Government (ICTBIG)*, 1-3. 10.1109/ICTBIG.2016.7892720

Shukla, S. N., & Champaneria, T. A. (2017). Survey of various data collection ways for smart transportation domain of smart city. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 681-685. 10.1109/I-SMAC.2017.8058265

Sun, Y., Song, H., Jara, H. J., & Bie, R. (2016). Internet of Things and Big Data Analytics for Smart and Connected Communities. *IEEE Access: Practical Innovations, Open Solutions*, 4, 766–773. doi:10.1109/ACCESS.2016.2529723

Zenkert, J., Dornhofer, M., Weber, C., Ngoukam, & Fathi, M. (2018). Big data analytics in smart mobility: Modeling and analysis of the Aarhus smart city dataset. *IEEE Industrial Cyber-Physical Systems (ICPS)*, 363-368. doi:10.1109/ICPHYS.2018.8387685

KEY TERMS AND DEFINITIONS

Big Data: Big data frequently refers to data sets with sizes away from the ability of frequently used software tools to confine, curate, administer, and process data contained by a supportable elapsed in-

stance. Big data philosophy encompasses shapeless, semi-structured and prepared data, however the main focal point is on unstructured data.

Big Data Analytics: Big data analytics is the often composite process of examining large and wide-ranging data sets, or big data, to come across information—such as concealed patterns, indefinite correlations, market development and client preferences—that can help organizations create informed business decisions.

Cloud Computing: Cloud computing is the on-demand ease of use of computer system resources, particularly data storage and computing influence, without direct active organization by the user. The expression is generally used to explain data centers obtainable to many users over the Internet.

Evolutionary Computing: In computer science, evolutionary computation is a ancestors of algorithms for comprehensive optimization inspired by biological evolution, and a coordinated section of artificial intelligence and soft computing using these algorithms.

Internet of Things: The Internet of Things (IoT) is a arrangement of organized computing devices such as mechanical equipment or digital machines, they can be objects or community that are provided with exclusive identifiers and the facility to transfer data over a set-up without requiring human-to-human or human-to-computer communication.

Mobile Communication: Mobile communication allows broadcast of voice and multimedia data via a workstation or a mobile device exclusive of having connected to any physical or fixed connection. Mobile communication is growing day by day and has become a must have for every person. Mobile communication is the exchange of voice and data using a communication facility at the same time without any physical link.

Quantum Computing: All computing systems rely on an elementary ability to store and control information. Existing computers manipulate individual bits, which accumulate information as binary 0 and 1 states. Quantum computers focus more power on quantum mechanical phenomena to direct required information.

Data Streaming Processing Window Joined With Graphics Processing Units (GPUs)

Shen Lu

University of South Florida, USA

Richard S. Segall

Arkansas State University, USA

INTRODUCTION

A massive wave of changes in industries are created by data. Business models, applications and even data driven economy depend on data to overcome the stress of the competition and even ensure the certainty of the survival. Data can represent not only the development trends but also the values of the changes. With the time passing, the values of data decreases, which makes the usage of time-series time more important.

Big Data is Large-scale data and can be either discrete or continuous. This article entails research that discusses the continuous case of Big Data often called “Data Streaming”. More and more businesses will depend on being able to process and make decisions on streams of data. This article utilizes the algorithmic side of data stream processing often called “stream analytics” or “stream mining”.

In a streaming data system, the data is made available at the moment a client application needs it. As stated by Warski (2016) processing data in a streaming fashion has become more and more popular over the “traditional” way of batch processing of Big Data sets available as a whole. According to Warski (2016), the focus has shifted in the industry of “it’s no longer that important how big is your data but rather it’s much more important of how fast you can analyze it and gain insights.”

Data stream processing incorporates a notion of a Window on a stream as a way to convert an infinite stream into a finite relation in order to apply relational operations. The data stream can be based on either time or the tuple. Time-based execution provides a way to model simultaneity, and tuple-based execution provides a way to react to primitive events. To process the two different kinds of streams, one would need to find a convergence language in which the simple difference between the two streams would be resolved. For this data streaming Window Join, one would need to focus on time-based data stream.

BACKGROUND

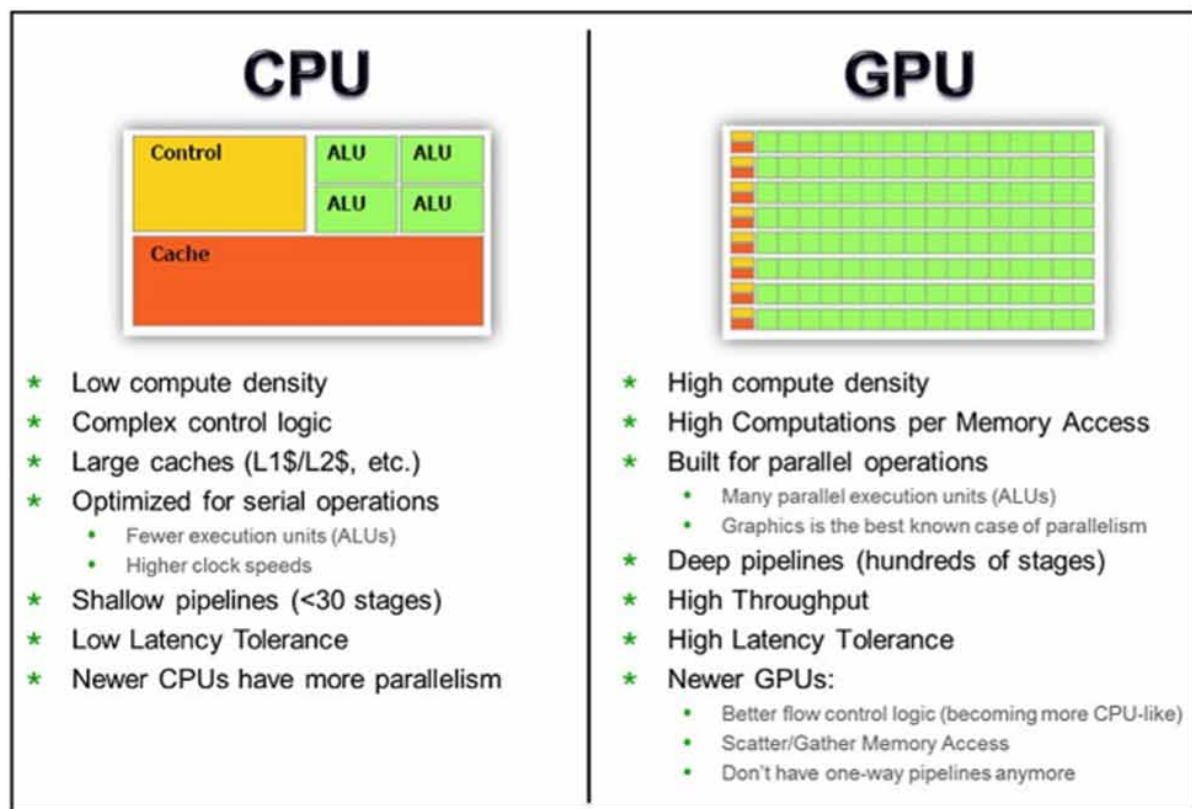
A Central Processing Unit (CPU) is the electronic circuitry within a computer that carries out the instructions of a computer program by performing the basic arithmetic, logical, control and input/output (I/O) operations specified by the instructions. The computer industry has used the term “central processing unit” at least since the early 1960s. Traditionally, the term “CPU” refers to a processor, more specifically to its processing unit and Control Unit (CU), distinguishing these core elements of a computer from external components such as main memory and I/O circuitry. (Wikipedia (2018e)).

DOI: 10.4018/978-1-7998-3479-3.ch043

A Graphics Processing Unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics and image processing, and their highly parallel structure makes them more efficient than general-purpose CPUs for algorithms where the processing of large blocks of data is done in parallel. In a personal computer, a GPU can be present on a video card, or it can be embedded on the motherboard or—in certain CPUs—on the CPU die. (Wikipedia (2018d)).

Figure 1 below shows the highlights of comparisons between Central Processing Unit (CPU) versus Graphics Processing Unit (GPU). Note that a CPU has a low compute density and a GPU has a high compute density. A CPU has a low latency tolerance and a GPU has a high latency tolerance, where latency refers to a measure of the time delay required for information to travel across a network. A CPU has shallow pipelines in contrast to a GPU that has high throughput. Hence this article discusses experiments using GPUs that are extremely more efficient than the traditional CPU framework.

Figure 1. Comparison of CPU (Central Processing Unit) vs. GPU (Graphics Processor Units (GPUs))



Below is a table of highlights of research advances from 2002 to 2018 in data streaming related to the data streaming processing of Window Joins with Graphics Processing Units (GPUs) and topic of processing algorithms for data analysis of streaming data.

Table 1. Highlights of Research Advances in Data Streaming (2018-2002)

YEAR	AUTHOR(S)	RESEARCH IN DATA STREAMING
2018	Akidau et al.	The What, Where, When, and How of Large-Scale Data Processing
2018	Aragues	Visualizing Data Streaming
2018	Bifet et. al.	Mining using frequent pattern, frequent itemset, and frequent subgraph.
2017	Psaltis	Data analysis of data streaming including sliding window and tumbling window techniques.
2017	Basak et. al.	Windowing for tumbling windows, hopping windows, and sliding windows; JOIN command for integrating streaming data for temporarily constrained data sources.
2016	Dunning & Friedman	Stream architecture using Apache Kafka and MapR; fraud detection with streaming data and geo-distributed data streams.
2016	Galic	Spatio-temporal data streams with time-based windows, tuple-based windows, predictive-based windows, and data stream clustering.
2016	Garofalakis et. al.	Data-stream sampling and sliding-window computation model.
2015	Dallachiesa et. al.	Showed sort-based similarity join algorithms can perform better than index-based techniques.
2015	Pinnecke et. al.	Splitting of variable-length windows into fixed size for continuous processing of data streams.
2014	Ellis	Sliding window reservoir sampling.
2013	Kim & Kim	Review of Window Query Processing for Data Streams
2011	Deshpande	Handbook for ORACLE Streams 11g.
2010	Bifet	Adaptive sliding algorithm ADWIN.
2010	Gama	Knowledge discovery in data streams using 31 algorithms
2007	Ghanem et al.	Supporting Views in Data Stream Management System (DSMS)
2006	Golab	Efficient query processing over sliding windows of Data Stream Management Systems.
2005	Govindaraju et. al.	Demonstrated Graphic Processing Units (GPUs) as efficient stream processor for mining data streams.
2003	Hammad et al.	Efficient execution of sliding-Window queries over data streams
2002	Babcock et. al.	Investigated stream query languages and processing, and algorithmic issues.

2018

Akidau et al. (2018) presents and discusses many figures for pattern analysis for data streaming and windowing strategies (See Figure 2 below) including that for variations for fixed windows, sliding windows, and sessions. Akidau et al. (2018) also discusses event-time windowing and processing-time windows using triggers and ingress times.

Aragues (2018) discusses in-depth chapters on processing streaming data for visualization and streaming visualization techniques, and also how to show machine learning results (such as scores or classifications) as data is classified in a streaming workflow. Aragues (2018) also states that visualizations can help us understand security using the relationships between and timelines of various threats, as security data often requires secondary lookups for history and context that can also be shown in a streaming visualization.

Bifet et al. (2018) is another book on subject of machine learning for data streams with practical examples in Massive Online Analysis (MOA) that includes discussions mining of streams using (1.) frequent pattern, (2.) frequent itemset and (3.) frequent subgraph.

2017

Basak et al. (2017) discusses stream analytics with Microsoft Azure and also the JOIN command for integrating streaming data that requires the data sources to be temporally constrained. Azure Stream Analytics is a fully managed cost effective real-time event-processing engine that helps to unlock deep insights from data. Stream Analytics makes it easy to set up real-time analytic computations on data streaming from devices, sensors, web sites, social media, applications, infrastructure systems, and more. Basak et al. (2017) also discusses windowing for tumbling windows, hopping windows, and sliding windows.

Psaltis (2017) wrote an entire book on understanding the real-time pipeline of streaming data. Psaltis (2017) contains an entire chapter on algorithms for data analysis of streaming data including sliding window and tumbling window techniques and their support in popular stream-processing frameworks.

Familiar and Barnes (2017) discusses in depth real-time processing using Azure Storm Analytics using SQL query dialect features, and also security and identity authorization and authentication using threat modeling zones and IoT (Internet of Things).

2016

Dunning and Friedman (2016) authored a book on streaming architecture with new designs using Apache Kafka and MapR streams that included topics of fraud detection with streaming data and geo-distributed data streams. As Dunning and Friedman (2016) indicate MapR Streams has many differences with Apache Kafka that include running a much larger number of topics and applying policies such as time-to-live or controlled access to many topics as a group. The ability to set up a very large number of topics in MapR Streams lets the building of topics that reflect business goals rather than infrastructural limitations.

As Dunning and Friedman (2016) state: each MapR stream can contain hundreds of thousands or more topics and partitions, and each MapR cluster can have millions of streams. In comparison, an Apache Kafka cluster consisting of server processor called “brokers” each of which should not have more than a thousand partitions per broker.

Galic (2016) is a book on spatio-temporal data streams that discusses time-based windows, tuple-based windows, and predicate-based windows, and data stream clustering.

Garofalakis et al. (2016) edited a book on data stream management that includes individual chapters on topics on foundations and basic stream synopses such as data-stream sampling, distinct-values estimations over data streams, and the sliding-window computation model and results.

2015

Dallachiesa et al. (2015) studied uncertain data streams by showing how current-state-of-art techniques for answering similarity join techniques can be easily adopted to be used with these under a variety of configurations using real data. Dallachiesa et al. (2015) showed that their algorithms used to maintain uncertain sliding windows can efficiently operate while providing a high quality approximation in query answering. Dallachiesa et al. (2015) also showed that sort-based similarity join algorithms can perform better than index-based techniques when the number of possible values per tuple is low as in many real-world applications.

According to Pinnecke et al. (2015) stated that the need of continuous processing and analysis of data streams has increased rapidly and that very large windows cannot be passed directly to the GPU and proposed to split the variable-length windows into fixed sized windows. Pinnecke et al. (2015) presented

a time and space efficient, data-corruption flow concept for this task and identified open research challenges related to co-processing in the context of data stream processing.

2014

Ellis (2014) wrote a book on real-time analytics as techniques to analyze and visualize streaming data that including sampling from a streaming population such as “sliding window reservoir sampling”.

2013

Kim & Kim (2013) provided a review of Window query processing for data streams that includes discussions of query plan structure, blocking operators, operator scheduling, load shedding and disorder control. Load shedding algorithms for Window joins can be classified into frequency-based model, age-based model, and pattern-based model. A data stream is an ordered sequence of data items, and the order is important in windowing data streams.

2011

Deshpande (2011) wrote a book on designing and managing a powerful data replication solution for the software product of ORACLE Streams 11g that depicts the Streams information flow as consisting of the three stages of: capture, staging and propagation, and consumption.

2010

Bifet (2010) is book that focuses on the design of learning algorithms for evolving and time-changing data streams called “adaptive streaming”. The basic building block in Bifet (2010) is the adaptive sliding algorithm, ADWIN, for change detection and value estimation. Bifet (2010) makes significant contributions to mining time-changing streams.

Gama (2010) discussed knowledge discovery from data streams using thirty-one different algorithms, novelty detection in data streams, and time-series data streams.

2007

Ghanem et al. (2007) proposed the Synchronized SQL (or SyncSQL) query language that defines a stream as a sequence of modify operations (i.e. insert, update, and delete) against relations with a specified schema. As Ghanem et al. (2007) indicated an important issue in continuous queries over data streams is the frequency by which the answer gets refreshed and the conditions that trigger the refresh. Coarser periodic refresh requirements are typically expressed as sliding-window queries.

2006

Sliding window query processing over data streams has been studied by Golab (2006) who developed solutions for efficient query processing over sliding windows of Data Stream Management Systems (DSMS) where there existed opportunities for resource sharing among similar queries because a large number of persistent queries could be executed at the same time.

2005

Previous related work utilizing Graphics Processing Units (GPUs) include that of Govindaraju et al. (2005) who presented algorithms for fast quantile and frequency estimation in large data streams. Govindaraju et al. (2005) used GPU as a co-processor to minimize the data transmission between CPU (Central Processing Unit) and GPU by taking into account the low bus bandwidth with results that demonstrate that the graphics processor available on a commodity computer system is an efficient stream-processor and a useful co-processor for mining data streams.

2003

Hammad et al. (2003) showed how the straightforward application of traditional pipelined query processing techniques to sliding Windows queries could result in inefficient and incorrect behavior. Hammad et al. (2003) presented three alternative execution techniques that guarantee correct behavior for pipelined sliding window queries and develop new algorithms for correctly evaluating Window-based duplicate-elimination, Group-By and Set operations in this context.

2002

Babcock et al. (2002) researched models and issues in data stream systems as models of multiple, continuous, rapid, time varying systems. Babcock et al. (2002) explored topics in stream query languages, new requirements and challenges in query processing, and algorithmic issues.

DATA STREAMING WINDOWS JOIN

Xie and Yang (2007) presented a survey of join processing in data streams. Datastreams.com (2018) discussed different ways of joining streams using Apache Spark 2.3.0 for stream-stream joins in Structured Streaming. Das et al. (2003) considered the problem of approximating sliding window joins over data streams in a data stream processing system with limited resources. Golub and Ozsu (2003) studies processing window multi-joins in continuous queries over data streams.

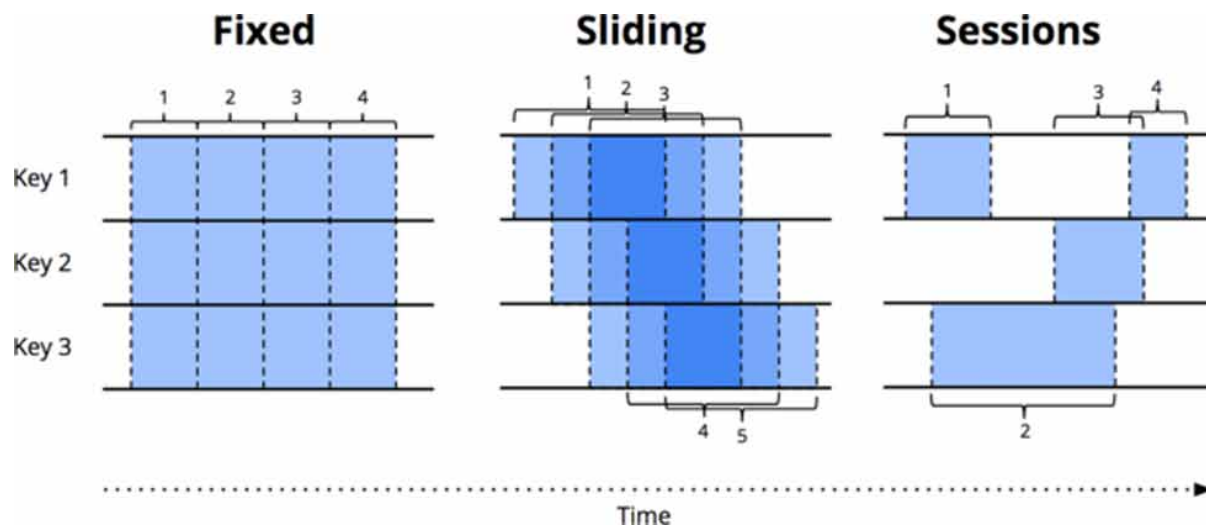
There are several types of Windowing strategies, and Akidau et al. (2018) illustrate and discuss the three types of (1.) fixed, (2.) sliding, and (3.) sessions as illustrated in the below Figure 2.

Using Graphics Processing Unit (GPU) can improve data streaming Window Joins for higher performance computing. Two independent threads generate data streams: one thread can be used to generate data stream A, and the other thread can be used to generate data stream B.

One would use a Windows Join thread to merge the two data streams, which is also the process of “Data Stream Window Join” as used in the title of this article. The Window Join Process can be implemented in parallel that can efficiently improve the computing speed. Data streams are loaded into multiple-thread GPU (Graphic Processing Unit) card, each data unit is loaded into a thread, and all of the data units can do computing at the same time, which can efficiently improve the throughput.

Because GPU (Graphics Processing Unit) machine has global memory and GPU card as central computing unit to store and process data, one would need special data structures to store data streams, array pointers, and temporary variables.

Figure 2. Windowing Strategies: Each example is shown for three different keys, highlighting the difference between aligned windows and unaligned windows. (Source: Akidau et al. (2018), p.17 and 33).



GPU has multiple-threading feature that makes the computing process different from single thread computing. To take advantage of this feature, one would need to put computing units into as many threads as possible. This article discusses choosing hash-table as the data structure to sort and store data, because, it can make the best use of the multiple-threading feature and efficiently improve the computing performance.

The `cudaMemcpyToSymbol` can be used to copy bytes from the memory area pointed to by source to the memory area pointed to by offset bytes from the start of symbol. The memory areas may not overlap. The symbol can either be a variable that resides in global or constant memory space, or it can be a character string, naming a variable that resides in global or constant memory space. As CUDA Programming Blogspot (2013) indicates, we use “constant memory” for data that will not change over the course of a kernel execution.

The kind can be either `cudaMemcpyHostToDevice` or `cudaMemcpyDeviceToDevice`. We use `cudaMemcpyToSymbol` to assign a global memory space to a circle queue and keep it in global memory as long as the data streams are going from host machine to device. It can be updated when the sliding window moves from time to time. We check the out of day of the time stamp from the front to the end. If the time stamp of the data item is out of date, the data can be deleted. We insert the data to the end of the queue. We need to check if the queue is full or not. Technically, if the GPU computing speed is faster than or equal to the CPU data generating speed, the circle queue cannot get full.

For GPU Data Stream Window Join, there are two steps: one is hash two data streams into two different arrays; the other is merge the two arrays and print the results.

PSEUDOCODE

According to Wikipedia (2018), a Pseudocode is an informal high-level description of the operating principle of a computer program or other algorithm. It uses the structural conventions of a normal programming language, but is intended for human reading rather than machine reading. That is, a Pseudocode is an artificial and informal language that helps programmers develop algorithms.

The “Pseudocode” provided in the Appendix of this article is composed of two threads: one is for consumer, the other is for producer. When the program is loaded, one of the threads is executed continuously without termination. Producer thread is then used to generate data following the use of consumer thread to do Windows Join. Producer function creates two streams. Consumer function is then used on the two streams to build and update the histogram, and generate results.

The use of Graphics Processing Unit (GPU) offers the high throughput of batch processors, the low latency of stream processors, and the ability to perform iterative and incremental computations. High-level programming models can be built on low-level primitives enabling such diverse tasks as streaming data analysis, iterative machine learning, and iterative graph mining.

CUDA® is a parallel computing platform and programming model developed by NVIDIA for general computing on Graphics Processing Units (GPUs). With CUDA, developers are able to dramatically speed up computing applications by harnessing the power of GPUs. For Windows Join on a CUDA machine, there are three steps:

1. Distribute incoming data to the hash table.
2. Update hash table: delete out-of-date entities.
3. Merge two hash tables.

These three steps take advantage of the multiple processors in GPU in order to gain high throughput. Both hash table data structure and the data merge are based on Parallel Computing. Data are distributed on difference threads, each thread can work independently to map to send data to corresponding hash bucket and select data from the other hash table to merge and to generate results. All of the threads can work simultaneously which can ensure the improvement of the throughput twenty to thirty times faster.

EXPERIMENTS

The data streams were generated on multiple-thread single-processor Windows workstation, as shown in Table 2. The Window joins were performed on Linux server with GPU with the machine details as are listed in Table 3.

Table 2. Multiple-thread Single-processor Windows Workstation

	CPU
Architecture	X86_64
Processor	Intel(R) Core(TM) i7-3930K CPU @ 3.20GHz
Core(s) per socket	6
Socket	1
CPU MHz	1200.0
L1i cache	32k
L1d cache	32k
L2 cache	256k
L3 cache	12288k
Linux version	3.11.0-26-generic

Table 3. Linux Server with GPU (Graphics Processing Units)

	GPU
Product	GF110GL
Vendor	NVIDIA
Version	A1
Width	64 bits
Clock	33MHZ
capabilities	vga_controller bus_master cap_list rom
Configuration	Driver=NVIDIA Latency=0

Table 4. Throughput of Data Stream Window Join for different sizes of windows

Window Size (million seconds)	Time Cost (million seconds)
10	418.976
20	385.284
20	533.759
40	612.361
50	669.958
60	742.854
70	907.188
80	1020.676
90	1131.624
100	1254.414

We tested the hash join algorithm on a streaming Window from 10 million seconds to 100 million seconds. With the increase of the size of the streaming window, the throughput increase, as shown in Table 4 and Figure 3. The performance of the hash join on different data sets is shown in Tables 5 and 6 below. An individual thread handles each tuple. All of the data are processed in parallel.

We made two algorithms for Data Stream Window Join for the purpose of comparison: one is based on static data, the other is based on dynamic data. When we use static data to do window join, experimental results are shown in Table 5. The performance of the dynamic Data Stream Window Join is shown in Table 6.

For static Data Stream Window Join, we also measured the time cost of Window Join. For Dynamic Data Stream Window Join, the Window Join can happen any time when producing data stream. We cannot measure how long it took to join data streams. In Figure 3 is the time cost of the hash join on static data. With the increase of the Window size, the time cost is randomly changed. In Figure 4 is the time cost of the hash join on dynamic data. With the increase of the window size, the time cost increases.

As shown in Figure 4, using static data, the time in seconds for hash join in a Window is cyclical with global maximum using Window size of 45 followed with a global minimum using Window size of 50 and local minimum using Window size of 20.

As shown in Figure 5, using dynamic data, the time in seconds taken for hash join monotonically increases with increasing number of tuples in a Window. Using the dynamic data, there is a decrease of

Figure 3. Throughput for Data Stream Window Join

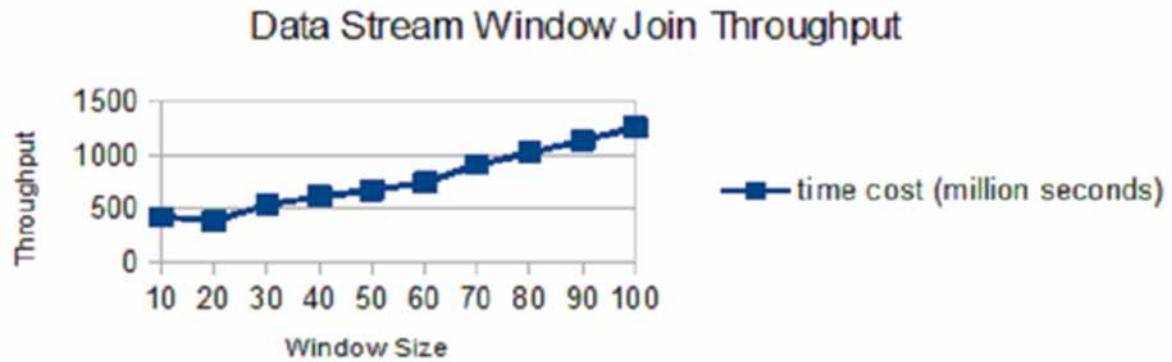


Table 5. Data Stream Window Join with Static Data

Window Size	Number of Tuples in R	Number of Tuples in S	Time taken for Hash Join
10	8214	8517	46.91
15	11011	11438	74.46
20	13165	13623	16.47
25	14847	15320	55.68
30	16224	16693	68.92
35	17344	17802	61.72
40	18285	18696	71.27
45	19038	19404	92.04
50	19712	20024	9.57
55	20245	20472	12.9

Table 6. Data Stream Window Join with Dynamic Data

Window Size	Number of Tuples in R	Number of Tuples in S	Time taken for Hash Join
10	14701	14701	10.7
15	23673	23673	22.92
20	38267	38267	39.73
25	27316	27316	82.51
30	39437	39437	117.65
35	54101	54101	148.21
40	61996	61996	189.46
45	90012	90012	244.47
50	75090	75090	354.47
55	77880	77880	514.9

Figure 4. Data Stream Window Join with Static Data

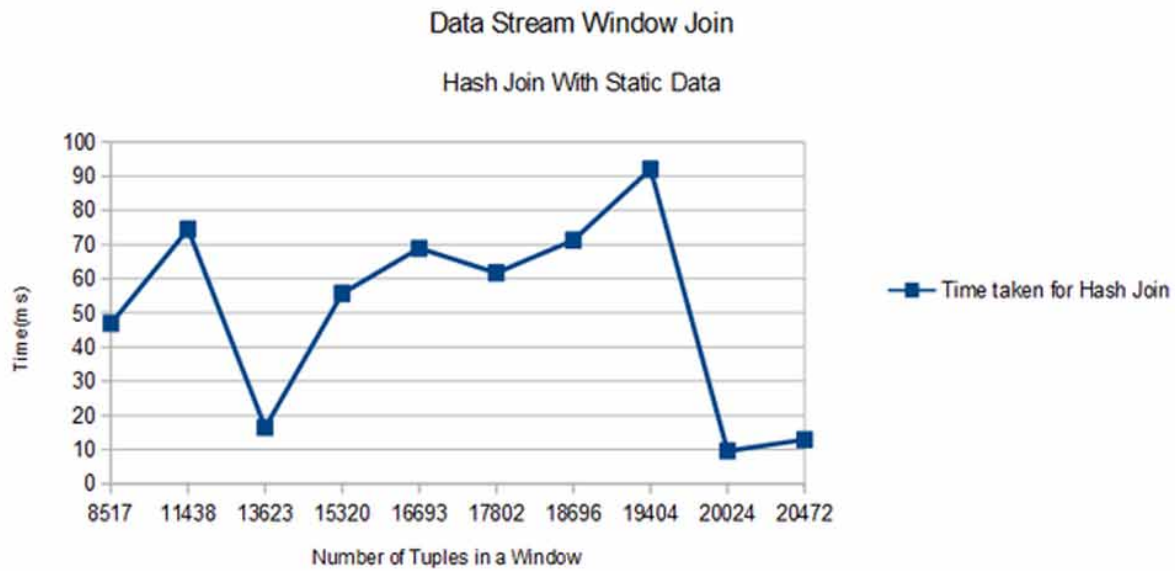
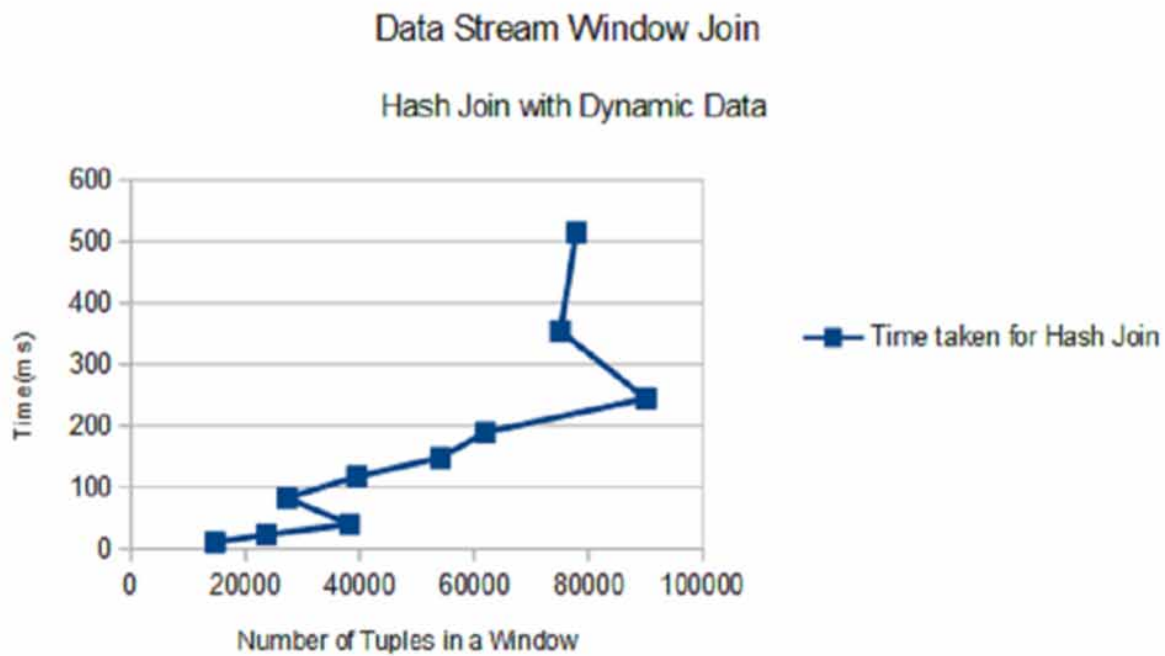


Figure 5. Data Stream Window Join with Dynamic Data



number of tuples in a Window when Window size is 25, and a maximum when the Window size is 45 that precedes a substantial decrease when Window size is 50 that precedes a marginal increase.

FUTURE RESEARCH DIRECTIONS

The future research directions are to perform additional experiments using both static and dynamic data of varying dimensionalities with different Graphics Processing Units (GPU) and make corresponding comparisons, as well as investigating the relationship between stream processing and General Purpose GPUs (GPGPU) as discussed below.

With the emergence of deep learning, the importance of GPUs has increased. In research done by Indigo, it was found that while training deep learning neural networks, GPUs can be 250 times faster than CPUs. The explosive growth of Deep Learning in recent years has been attributed to the emergence of General Purpose GPUs. There has been some level of competition in this area with Application-Specific Integrated Circuits (ASIC), most prominently the Tensor Processing Unit (TPU) made by Google. However, these can require changes to existing code and GPUs are still very popular. (Wikipedia (2018d))

Stream Processing and General Purpose GPUs (GPGPU)

It is becoming increasingly common to use a General Purpose Graphics Processing Unit (GPGPU) as a modified form of stream processor (or a vector processor), running compute kernels. This concept turns the massive computational power of a modern graphics accelerator's shader pipeline into general-purpose computing power, as opposed to being hard-wired solely to do graphical operations. In certain applications requiring massive vector operations, this can yield several orders of magnitude higher performance than a conventional CPU. (Wikipedia (2018d)).

The two largest discrete GPU designers, AMD (Advanced Micro Devices, Inc.) and Nvidia, are beginning to pursue this approach with an array of applications. Both Nvidia and AMD have teamed with Stanford University to create a GPU-based client for the Folding@home distributed computing project, for protein folding calculations. In certain circumstances, the GPU calculates forty times faster than the conventional CPUs traditionally used by such applications. (Wikipedia (2018d)).

GPGPU can be used for many types of embarrassingly parallel (Wikipedia (2018f)) tasks including ray tracing. They are generally suited to high-throughput type computations that exhibit data-parallelism to exploit the wide vector width Single Instruction Multiple Data (SIMD) architecture of the GPU.

Furthermore, GPU-based high performance computers are starting to play a significant role in large-scale modelling. According to the Top500 (2012) supercomputers list, three of the 10 most powerful supercomputers in the world took advantage of GPU acceleration.

CONCLUSION

This article compared the differences of Central Processing Units (CPU) versus Graphics Processing Units (GPU), discussed and summarized advances in data streaming and its relationship to data streaming processing of Window Joins with Graphics Processor Units (GPUs), and compared processing of data streaming with experiments performed using a Pseudocode as provided in Appendix.

The comparisons of data stream Window Joins are made with both static and dynamic data. The article illustrates the endless potential of the importance and applicability of data streaming for processing using technology such as Graphics Processing Units.

REFERENCES

- Akidau, T., Chernyak, S., & Lax, R. (2018). *Streaming Systems: The What, Where, and How of Large-Scale Data Processing*. Retrieved August 18, 2018 from https://www.amazon.com/dp/1491983876/ref=emc_b_5_i
- Aragues, A. (2018). *Visualizing Streaming Data: Interactive Analysis beyond Static Limits*. Retrieved August 18, 2018 from https://www.amazon.com/dp/1492031852/ref=emc_b_5_i
- Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). *Models and Issues in Data Stream Systems*. Retrieved on May 16, 2018 from https://infolab.usc.edu/csci599/Fall2002/paper/DML2_streams-issues.pdf
- Basak, A., Venkatraman, K., Murphy, R., & Singh, M. (2017). *Stream Analytics with Microsoft Azure: Real-time data processing for quick insights using Azure Stream Analytics*. Retrieved on May 16, 2018 from https://www.amazon.com/gp/product/1788395905/ref=oh_aui_detailpage_o03_s00?ie=UTF8&psc=1
- Bifet, A. (2010). *Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams*. IOS Press. The Netherlands, ISBN 978-1-60750-090-2. Retrieved July 16, 2020 from <https://www.iospress.nl/book/adaptive-stream-mining-pattern-learning-and-mining-from-evolving-data-streams/>
- Bifet, A., Gavalda, R., Holmes, G., & Pfahringer. (2018). *Machine Learning for Data Streams: with Practical Examples in MOA (Adaptive Computation and Machine Learning)*. MIT Press. Retrieved on May 18, 2018 from https://www.amazon.com/gp/product/0262037793/ref=oh_aui_detailpage_o03_s00?ie=UTF8&psc=1
- CUDA Programming BlogSpot. (2013). *What is “Constant Memory” in Cuda?* Blog posted by Gupta, N. Retrieved on June 3, 2018 from <http://cuda-programming.blogspot.com/2013/01/what-is-constant-memory-in-cuda.html>
- Dallachiesa, M., Jacques-Silav, G., Gedik, B., Wu, K.-L., & Palpanas, T. (2015). Sliding Windows over Uncertain Data. *Knowledge and Information Management*, 45(1), 159-190. Retrieved on May 13, 2018 from <https://link.springer.com/article/10.1007/s10115-014-0804-5>
- Das, A., Gehrke, J., & Riedewald, M. (2003). Approximate Join Processing over Data Streams. *ACM SIGMOD 2003*, 40-51. Retrieved on November 21, 2018 from <https://www.cs.cornell.edu/johannes/papers/2003/sigmod2003-loadShedding.pdf>
- Databricks.com. (2018). *Stream-Stream Joins using Structured Streaming (Scala)*. Retrieved November 21, 2018 from https://docs.databricks.com/_static/notebooks/stream-stream-joins-scala.html
- Deshpande, K. (2011). *Oracle Streams 11g Data Replication*, McGraw-Hill. Retrieved on May 18, 2018 from https://www.amazon.com/gp/product/0071496645/ref=oh_aui_detailpage_o04_s00?ie=UTF8&psc=1

Dunning, T., & Friedman, E. (2016). *Streaming Architecture: New Designs Using Apache Kafka and MapR Streams*. Retrieved on May 16, 2018 from https://www.amazon.com/gp/product/1491953926/ref=oh_aui_detailpage_o01_s01?ie=UTF8&psc=1

Ellis, B. (2014). *Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data*. John Wiley & Sons, Inc. Retrieved on May 16 2018 from https://www.amazon.com/gp/product/1118837916/ref=oh_aui_detailpage_o00_s00?ie=UTF8&psc=1

Familiar, B., & Barnes, J. (2017). *Business in Real-Time Using Azure IoT and Cortana Intelligence Suite: Driving your Digital Transformation*. Springer Science + Business Media.

Galic, Z. (2016). *Spatio-Temporal Data Streams*. SpringerBriefs in Computer Science. Retrieved on May 16, 2018 from https://www.amazon.com/gp/product/1493965735/ref=oh_aui_detailpage_o01_s00?ie=UTF8&psc=1

Gama, J. (2010). *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Retrieved on May 16, 2018 from https://www.amazon.com/gp/product/1439826110/ref=oh_aui_detailpage_o05_s00?ie=UTF8&psc=1

Garofalakis, M., & Gehrke, J. (2016). *Data Stream Management: Processing High-Speed Data Streams (Data-Centric Systems and Applications)*. Springer-Verlag. Retrieved on May 16, 2018 from https://www.amazon.com/gp/product/3540286071/ref=oh_aui_detailpage_o01_s01?ie=UTF8&psc=1

Ghanem, T. M., Aref, W. G., Elmagarmid, A. K., & Larson, P.-A. (2007). *Supporting Views in Data Stream Management System*. Purdue University e-Pubs, Department of Computer Science Technical Reports. Retrieved on November 20, 2018 from <https://docs.lib.purdue.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=2687&context=cstech>

Golab, L. (2004) Querying Sliding Windows over Online Data Streams. In *Current Trends in Database Technology - EDBT 2004 Workshops*. Springer. Retrieved on November 20, 2018 from https://link.springer.com/chapter/10.1007/978-3-540-30192-9_1

Golab, L. (2006). *Sliding Window Query Processing over Data Streams*. Technical Report CS-2006-27, Department of Computer Science, University of Waterloo. Retrieved on May 16, 2018 from <https://pdfs.semanticscholar.org/cafd/395e618f8a07341d9fc1d9ffbd8aeb0a3b8c.pdf>

Golab, L., & Ozsu, M. T. (2003). Processing Sliding Windows Multi-Joins in Continuous Queries over Data Streams. *Proceedings of the 29th International Conference on Very Large Databases (VLDB)*, 500-511. 10.1016/B978-012722442-8/50051-3

Govindaraju, N. K., Raghuvanshi, N., & Manocha, D. (2005). Fast and Approximate Stream Mining of Quantiles and Frequencies Using Graphic Processors. *SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data*, 611-622. Retrieved on May 16, 2018 from <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/10/13.pdf>

Hammad, M. A., Aref, W. G., Franklin, M. J., Mokbel, M. P., & Elmagarmid, A. K. (2003). Efficient Execution of Sliding-Windows Queries over Data Streams, *Purdue University e-Pubs, Department of Computer Science Technical Reports*. Retrieved on November 20, 2018 from <https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2583&context=cstech>

Jain, A. (2017). *Mastering Apache Storm: Real-time big data streaming using Kafka, Hbase and Redis*. Packt Publishing Ltd. Birmingham, UK. Retrieved July 16, 2020 from <https://www.packtpub.com/big-data-and-business-intelligence/mastering-apache-storm>

Kim, H. G., & Kim, M. H. (2013, December). A Review of Window Query Processing for Data Streams. *Journal of Computing Science and Engineering: JCSE*, 7(4), 220–230. doi:10.5626/JCSE.2013.7.4.220

Pinnecke, M., Broneske, D., & Saake, G. (2015). Toward GPU Accelerated Data Stream Processing. *27th GI-Workshop on Foundations of Databases*. Retrieved on May 16, 2018 from <http://ceur-ws.org/Vol-1366/paper15.pdf>

Psaltis, A. (2017). *Streaming Data: Understanding the real-time pipeline*. Manning Publications Co. Retrieved on May 16, 2018 from https://www.amazon.com/gp/product/1617292281/ref=oh_aui_detail_page_o07_s00?ie=UTF8&psc=1

Tejada, Z. (2017). *Mastering Azure Analytics: Architecting in the cloud with Azure Data Lake, HDInsight and Spark*. O'Reily Media Inc. Sebastopol, CA 95472. ISBN: 978-1-491-95665-6. Retrieved July 16, 2020 from <https://www.oreilly.com/library/view/mastering-azure-analytics/9781491956649/>

Top500 Org. (2012). *Top500 List*. Retrieved on August 26, 2018 from <https://www.top500.org/list/2012/06/100/>

Warski, A. (2016). *Windowing data in Big Data Streams - Spark, Flink, Kafka, Akka*, Retrieved on May 16, 2018 from <https://softwaremill.com/windowing-in-big-data-streams-spark-flink-kafka-akka/>

Wikipedia. (2018a). *Adaptive bitrate streaming*. Retrieved on June 7, 2018 from https://en.wikipedia.org/wiki/Adaptive_bitrate_streaming

Wikipedia. (2018b). *CUDA*. Retrieved on June 7, 2018 from <https://en.wikipedia.org/wiki/CUDA>

Wikipedia. (2018c). *Hash table*. Retrieved on June 7, 2018 from https://en.wikipedia.org/wiki/Hash_table

Wikipedia. (2018d). *Graphics Processing Unit (GPU)*. Retrieved on June 7, 2018 from https://en.wikipedia.org/wiki/Graphics_processing_unit

Wikipedia. (2018e). *Central Processing Unit (CPU)*. Retrieved on August 19, 2018 from https://en.wikipedia.org/wiki/Central_processing_unit

Wikipedia. (2018f). *Embarrassingly Parallel*. Retrieved on August 26, 2018 from https://en.wikipedia.org/wiki/Embarrassingly_parallel

Xie, J., & Yang, J. (2007). A Survey of Join Processing in Data Streams. In C. C. Aggarwal (Ed.), *Data Streams. Advances in Database Systems* (Vol. 31, pp. 209–236). Springer. . ISBN 978-0-387-28759-1. Retrieved July 16, 2020 from https://link.springer.com/chapter/10.1007/978-0-387-47534-9_10 doi:10.1007/978-0-387-47534-9_10

ADDITIONAL READING

Goebel, V. (n.d.), *Data Stream Management Systems: Applications, Concepts and Systems*, University of Oslo, PowerPoint presentation retrieved on November 20, 2018 from <https://www.uio.no/studier/emner/matnat/ifi/INF5100/h06/undervisningsmateriale/handouts/DSMS.pdf>

Gottardo, M. (2018). *FPGA to High speed ADC Data streaming, HDL programming: Xilinx Zynq7000 family on Vivado IDE platform (FPGA and SoC programming)*. Vigonova, Italy. Retrieved August 18, 2018 from https://www.amazon.com/FPGA-High-speed-streaming-programming/dp/1720843694/ref=sr_1_15?s=books&ie=UTF8&qid=1534644932&sr=1-15&keywords=data+streaming

Karim, M. R., & Alla, S. (2018). *Scala and Spark for Big Data Analytics: Explore the concepts of functional programming, data streaming, and machine learning*, Packt Publishing. Birmingham, UK. ISBN-13: 978-1785280849. Retrieved August 18, 2018 from https://www.amazon.com/Scala-Spark-Big-Data-Analytics/dp/1785280848/ref=sr_1_11?s=books&ie=UTF8&qid=1534644932&sr=1-11&keywords=data+streaming

Kumar, M., & Singh, C. (2018). *Building Data Streaming Applications with Apache Kafka: Design, develop and streamline applications using Apache Kafka, Storm, Heron and Spark*, Packt Publishing, Birmingham, UK. ISBN-13: 978-1787283985. Retrieved August 18, 2018 from https://www.amazon.com/Building-Streaming-Applications-Apache-Kafka/dp/1787283984/ref=sr_1_5?s=books&ie=UTF8&qid=1534644932&sr=1-5&keywords=data+streaming

Modi, R. (2019). *Azure for Architects: Implementing cloud design, DevOps, IoT, and serverless solutions on your public cloud*. Second Edition. Packt Publishing Ltd., Birmingham, UK. Retrieved July 16, 2020 from <https://www.packtpub.com/virtualization-and-cloud/azure-architects-second-edition>.

Saxena, S., & Gupta, S. (2017). *Practical Real-Time Data Processing and Analytics*. Packt Publishing Ltd., Birmingham, UK. Retrieved July 16, 2020 from <https://www.packtpub.com/big-data-and-business-intelligence/practical-real-time-data-processing-and-analytics>.

Smith, M. D. (2017). *Streaming, Sharing, Stealing: Big Data and the Future of Entertainment*, The MIT Press, Cambridge, MA. 1st Edition, ISBN-13: 978-0262034791. Retrieved August 18, 2018 from https://www.amazon.com/Streaming-Sharing-Stealing-Future-Entertainment/dp/0262034794/ref=sr_1_7?s=books&ie=UTF8&qid=1534644932&sr=1-7&keywords=data+streaming

Zadrozny, P. (2013). *Big Data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources*, Apress Publishers, New York, NY. ISBN-13: 978-1430257615. Retrieved August 18, 2018 from https://www.amazon.com/Big-Data-Analytics-Using-Splunk/dp/143025761X/ref=sr_1_24?s=books&ie=UTF8&qid=1534650233&sr=1-24&keywords=data+streaming

Zikopoulos, P. C. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, Data, New York, NY. ISBN-13: 978-0071790536. Retrieved August 18, 2018 from https://www.amazon.com/Understanding-Big-Data-Analytics-Enterprise/dp/0071790535/ref=sr_1_20?s=books&ie=UTF8&qid=1534650114&sr=1-20&keywords=data+streaming

KEY TERMS AND DEFINITIONS

Adaptive Streaming: Technique used in streaming multimedia over computer networks that works by detecting a user's bandwidth and CPU capacity in real time and adjusting the quality of the media stream accordingly (Wikipedia, 2018a).

ADWIN: Adaptive sliding window algorithm for change detection and value estimation with strong theoretical guarantees (Gama, 2010).

CUDA: Is a parallel computing platform and application programming interface (API) model created by Nvidia. It allows software developers and software engineers to use a CUDA-enabled Graphics Processing Unit (GPU) for general purpose processing – an approach termed GPGPU (General-Purpose computing on Graphics Processing Units; Wikipedia, 2018b).

cudaMemcpyToSymbol: Used to copy bytes from the memory area pointed to by source to the memory area pointed to by offset bytes from the start of symbol.

Data Stream: Continuous unbounded sequence of data values that arrive in time at a port on a system (Govindaraju et al., 2005, p. 613).

Data Stream Management System (DSMS): Database management system (DBMS) adopted to the continuous unbounded sequence of data values that compose data stream (s) (Babcock et al., 2002, p. 1).

Embarrassingly Parallel: Workload or problem (also called perfectly parallel or pleasingly parallel) is one where little or no effort is needed to separate the problem into a number of parallel tasks. This is often the case where there is little or no dependency or need for communication between those parallel tasks, or for results between them (Wikipedia, 2018f).

Graphics Processing Unit (GPU): A specialized computer processor, addresses the demands of real-time high-resolution 3D graphics compute-intensive tasks (Wikipedia, 2018d).

Hash Table (Hash Map): A data structure which implements an associative array abstract data type, a structure that can map keys to values. A hash table uses a hash function to compute an index into an array of buckets or slots, from which the desired value can be found (Wikipedia, 2018c).

Hopping Windows: Group events for aggregation and are windows of equal duration but overlap at regular intervals (Basak et al., 2017, p. 115).

Sliding Windows: Define all possible intervals of equal duration and output aggregations based on actual event stream activity (Basak et al., 2017, p. 116).

Streaming Data System: A non-hard real-time system that makes its data available at the moment a client application needs it (Psaltis, 2017, pp. 8-9).

Streaming Mining: Knowledge discovery from data streams (Gama, 2010).

Tumbling Windows: Intervals defined to group events for aggregation, and are windows that are contiguous and of equal duration (Basak et al., 2017, p. 113).

Windows Join: Integrates two or more inputs or input-derived data sets, just as in SQL (Structured Query Language) but requires the data sources to be temporarily constrained to match the sources for a window of time (Basak et al., 2017, p.111).

APPENDIX

5

Pseudocode Used in Experiments

```

void *producer(void *ptr){

while(true){

pthread_mutex_lock(&the_mutex);

while(front%MAX==(end%MAX-1)) pthread_cond_wait(&cond, &the_mutex);

/*

1. Allocate memory

2. generate data

3. copy to global memory

*/

time_t timestamp_sec; /* timestamp in second */

time(&timestamp_sec); /* get current time; same as: timestamp_sec = time(NULL) */

atom_list[front%MAX].key = front%MAX;

atom_list[front%MAX].value = front%MAX;

atom_list[front%MAX].timestamp = timestamp_sec;

cudaStatus = cudaMemcpy(dev_atom_list, atom_list, sizeof(atom), cudaMemcpyHostToDevice);

if(cudaStatus != cudaSuccess){

fprintf(stderr, "cudaMemcpy failed!");

exit(0);

}

```

```
printf("producer %i\n", front%MAX);

front++;

pthread_cond_signal(&cond);

pthread_mutex_unlock(&the_mutex);

}

pthread_exit(0);

}

void *consumer(void *ptr){

while(true){

pthread_mutex_lock(&the_mutexR);

pthread_mutex_lock(&the_mutexS);

while(endR%MAX==(frontR%MAX-1)) pthread_cond_wait(&cond, &the_mutexR);

while(endS%MAX==(frontS%MAX-1)) pthread_cond_wait(&cond, &the_mutexS);

/*

1. get data from global memory

2. compare, join, and update data

*/

int * d_end;

cudaMalloc((void**)&d_end, sizeof(int));

cudaMemcpy(d_end, &end, sizeof(int), cudaMemcpyHostToDevice);

windowJoin<<<grids,threads>>>(dev_atom_R, dev_atom_S, frontR, endR, frontS, endS, d_joinCount,
d_endR, d_endS);

cudaDeviceSynchronize();
```

```

cudaMemcpy(&output, d_joinCount, sizeof(int), cudaMemcpyDeviceToHost);

cudaMemcpy(&end, d_end, sizeof(int), cudaMemcpyDeviceToHost);

pthread_cond_signal(&cond);

pthread_mutex_unlock(&the_mutexR);

pthread_mutex_unlock(&the_mutexS);

}

pthread_exit(0);

}

//__global__ void windowJoin(int* atom_list, int front, int end, int* d_joinCount, int* d_updatedEnd){

__global__ void windowJoin(atom* atomR, atom* atomS, int frontR, int endR, int frontS, int endS, int*
d_joinCount, int* d_endR, int* d_endS){

insertData(atom_listR, atomR, endR)

insertData(atom_listS, atomS, endS)

// 1. build histogram

Histo(atom_listR, atomR, frontR, endR, d_endR, dev_histR, dev_atom_listR);

Histo(atom_listS, atomS, frontS, endS, d_endS, dev_histS, dev_atom_listS);

// 2. summarize the number joined records from R stream and S stream

probe(dev_histR, dev_atom_listR, dev_histS, dev_atom_listS, d_joinCount);

}

atom * atom_list;

atom * dev_atom_R, dev_atom_S;

cudaError_t cudaStatus;

pthread_mutex_t the_mutexR, the_mutex_S;

```

```
pthread_cond_t condR, condS;
```

Function

Input: int deviceID, windowSize, RstreamSize, SstreamSize

```
{  
  
//int dev = atoi(argv[1]);  
  
//plan.context = CreateCudaDevice(dev);  
  
//plan.context2 = CreateCudaDeviceStream(dev);  
  
start_time = clock();  
  
front = 0;  
  
end = 0;  
  
atom_list = (atom *)malloc(sizeof(atom)*MAX);  
  
cudaStatus = cudaSetDevice(0);  
  
cudaStatus = cudaMalloc((void**)&dev_atom_list, MAX*sizeof(atom));  
  
if(cudaStatus != cudaSuccess){  
  
fprintf(stderr, "cudaMalloc failed!");  
  
exit(0);  
  
}  
  
pthread_t pro, con;  
  
pthread_mutex_init(&the_mutexR, 0);  
  
pthread_mutex_init(&the_mutexS, 0);  
  
pthread_cond_init(&condR, 0);  
  
pthread_cond_init(&condS, 0);  
  
pthread_create(&con, 0, consumer, 0);
```

```
pthread_create(&pro, 0, producerR, 0);  
pthread_create(&pro, 0, producerS, 0);  
pthread_join(pro, 0);  
pthread_join(con, 0);  
pthread_cond_destroy(&condR);  
pthread_cond_destroy(&condS);  
pthread_mutex_destroy(&the_mutexR);  
pthread_mutex_destroy(&the_mutexS);  
return 0;  
}
```

Statistical Techniques for Research

Jose Carlos Casas-Rosal

 <https://orcid.org/0000-0002-6954-6517>

Universidad de Córdoba, Spain

Carmen León-Mantero

 <https://orcid.org/0000-0002-2870-8807>

Universidad de Córdoba, Spain

Noelia Jiménez-Fanjul

 <https://orcid.org/0000-0002-5728-8725>

Universidad de Córdoba, Spain

Alexander Maz-Machado

 <https://orcid.org/0000-0002-4112-4363>

Universidad de Córdoba, Spain

INTRODUCTION

Data analysis of a study and its subsequent statistical processing represent the base of numerous research works carried out in any scientific field. The wide variety of techniques allow an in-depth analysis of any data set depending on the desired goal – such goal has to be taken into account at the beginning of the study and it has to assume central importance during the research design, as it requires the verification of a set of hypotheses. The refutation of one or more hypothesis needed for the application of a particular statistical technique may lead to inconsistent or invalid results, which would in turn invalidate the whole research.

The vast majority of techniques applied in research is aimed at inferring the results of a sample to the target population of the study, which at the same time entails a certain degree of uncertainty – this degree of uncertainty can be controllable and measurable as long as the application conditions of the techniques are met. Its refutation may involve from a slight increase of the degree of uncertainty to its immeasurability. It is crucial for the researcher to possess a proper knowledge of the technique so as to obtain the results and the conclusions of his investigation.

However, and due to the great variety of techniques, its misuse in research is very common, either because of the refutation of any of the hypothesis set or because of the nature of the data that is being analysed. Because of that, the results obtained from its application cannot be taken into consideration due to its lack of validity.

Some of the main statistical techniques used in different research fields are presented in this chapter. A brief definition of each technique, together with the specification of its utility is included as well, in order to determine accurately its purpose. The required hypothesis regarding data and regarding experimental design are listed as well. Besides, some of the computer programs with which these techniques can be used are listed – we emphasize in R, a free and open-source software with multiple packages that allow to apply these techniques in an effective and simple way.

DOI: 10.4018/978-1-7998-3479-3.ch044

The analysis of the great amount of statistical techniques that exist would require a much larger extension than that available in this chapter, so the most used ones have been selected. Some of the simplest ones have been included (such as descriptive analysis, graphic analysis or parametric and non-parametric hypothesis testing for statistical inference), as well as some multivariate analysis techniques, such as principal component analysis, factor analysis, structural equation models, regression analysis and causal models.

BACKGROUND

The word *Statistics* comes from the Italian word *Statista*, meaning “Statesman” – referring to its first meaning as a science used by the Government to collect socio-demographic information from its population. From its introduction, it has experimented several extensions in its definition. During the nineteenth century, its field of action was enlarged to the “recollection of information from any data set”, and therefore, its application to other disciplines was extended as well. Later, the analysis and interpretation of data was also considered as a part of the discipline.

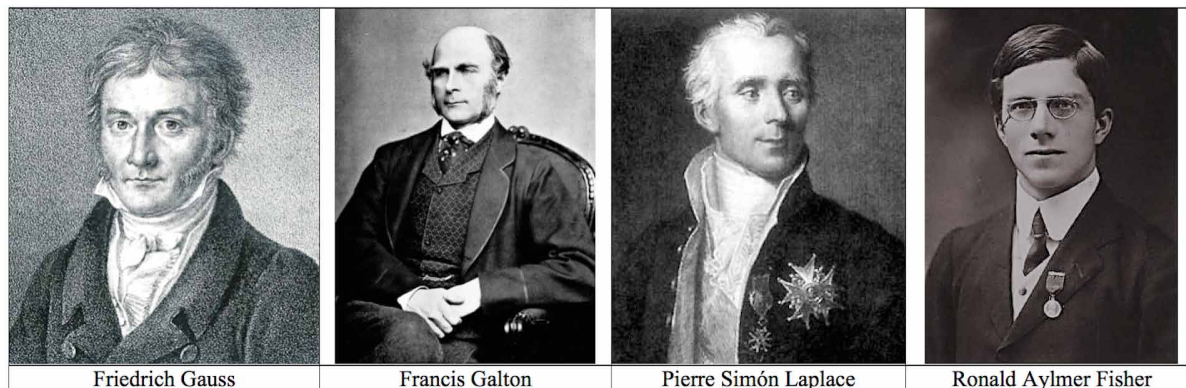
However, even though the origin of the word dates from the eighteenth century, Statistics had already been used by the Babylonian and Egyptian civilizations in the elaboration of frequency tables and population censuses. Several authors consider four fundamental periods during the development of Statistics: a period up to 1750 is characterized by the development of probability and the exposition of non-probabilistic methods of data analysis; a period from 1750 to 1820 when inference and mathematical statistics were introduced with works from Laplace and Gauss; a period from 1820 to the early twentieth century with works from Galton, Pearson and Fisher, when statistical inference, correlation and statistical models were developed; and a fourth period, during the last third of the twentieth century, with a strong evolution of Computer Science (Fienberg, 1992). Since then, the amount of techniques developed has been significantly extended due to the calculation capacity of computers.

Therefore, descriptive analysis and the elaboration of representative graphics can be framed within the beginning of Statistics as a discipline, while regression analysis and statistical inference were developed during the nineteenth century, whose most relevant researchers in the field are shown in Figure 1. Other techniques were developed during the twentieth century. And most of those were developed during its second half, after the revolution that took place due to the calculation capacity of computers. Factor analysis was developed by Charles Spearman during 1904 and later on by Louis Leon Thurstone. Due to the evolution of this methodology, together with Biometrics, Econometrics and Computing, structural equation models were born during the 70's. Principal component analysis is due to Hotelling (1933). And by the end of the twentieth century, artificial neural networks started to be applied to numerous statistical problems such as estimation or classification.

This great development has favoured the appearance of a large number of research works that use statistics to obtain important results that permit correct decision-making in important fields such as Health Sciences, Engineering or Social Sciences. Kaplan-Meier estimator, for instance, was developed in 1958 (Kaplan & Meier, 1958) and has been used in Engineering, Medicine and Economy to estimate the survival of the components of a machine, the survival of patients of a certain disease or the time it takes for a person to return to the active market after losing his job.

Each and every one of the existing statistical techniques are based on a mathematical basis that allows to demarcate with exactitude the conditions of applicability of the technique and its degree of uncertainty, if any. The application of a technique under different conditions of those for which it was designed is

Figure 1. Most relevant researchers in Statistics during 19th century

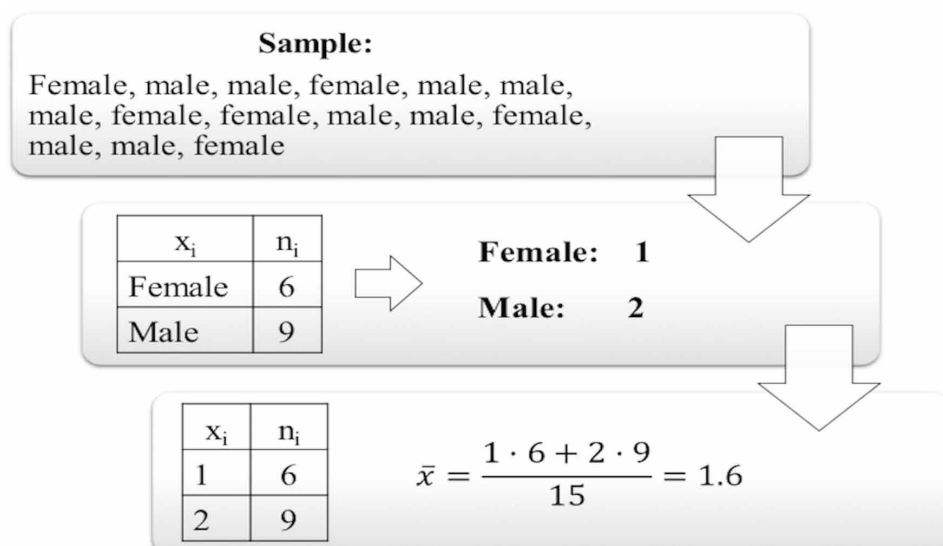


nonsense, although the technique may be applied numerically. To illustrate the previous statement, let X be the variable *Gender* measured for a group of individuals belonging to a certain population that contains the categories x_1 : “Man” and x_2 : “Woman”. The impossibility of calculating the mean of all values is obvious, since its mathematical expression would be:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2}{n_1 + n_2}$$

Where n_1 and n_2 are the number of men and women, respectively, of the study sample. Its categories are not numeric, so the calculation is not possible. On the contrary, the calculation would be possible if numerical values were assigned to each category, for example, 1 to x_1 : “female” and 2 to x_2 : “male”, so the result would be a value comprised between 1 and 2 that, even though it can be calculated numerically, is nonsense (Figure 2).

Figure 2. Example of the arithmetic mean for a dichotomous variable



Only in the case that the values 0 and 1 were assigned to the two modalities, there would be a coincidence between the value of the arithmetic mean and the proportion of individuals that verify modality 1, and therefore its calculation would make sense. In the example of Figure 2, if the value 1 is assigned to female and 0 to male, the arithmetic mean would be:

$$\bar{x} = \frac{1 \cdot 6 + 0 \cdot 9}{15} = 0.4$$

Consequently, the arithmetic mean could be interpreted as the proportion of females in the analysed sample, that is, a 40% of the individuals, and also, a possible estimation of the population portion.

In order to carry out a proper research it is necessary to define, previously, the desired objective and the appropriate statistical techniques that could be applied to achieve such objective. Also, a correct definition of the target population must be made, as an unrealistic consideration of it can generate an important sampling bias. Because of that, it is necessary to be precise (Holmes, 2004).

Once this is done, it is time to start data collection for the subsequent implementation of the techniques. This process can be divided into two main parts: sample selection and the development of data collection instrument. If the study aims for the inference of the results of a sample to its population, it is necessary that - among many other characteristics that can be consulted in Lohr (2009) or Fowler Jr. (2013) - the sample selection process was random and representative of the population for which it was designed.

Besides, this data collection instrument must be valid and consistent with the research objective. The most used data collection instrument is the *opinion questionnaire*, for which the following stages can be established for its elaboration: (1) the type of information needed is clearly defined, as well as the individuals for whom the instrument is being prepared; (2) the type of questions that are to be included (open or closed) are determined and written according to principles of simplicity, clarity and objectivity; (3) an introductory text is written and the instructions for the interviewer and the respondent are elaborated; and (4) the formal design of the questionnaire is included in the final stage of the elaboration of the questionnaire. Once constructed, the reliability and validity of the questionnaire must be analysed, as well as the suitability of the variables obtained through it, so as to verify the proper application of the desired statistical techniques. This can be consulted with greater detail in Blair, Czaja and Blair (2013). Through this instrument the variables describing the available information can be obtained - they must be correctly defined so that their modalities are exclusive. A variable can be qualitative or quantitative depending on the way in which the information is collected. This must be chosen, once more, according to the objective pursued. For example, the variable *Age* can be measured discretely, by intervals or qualitatively by means of a categorization, if the desire is to determine if the individual verifies a condition related to it. As we have discussed previously, the quantitative or qualitative nature of the variable determines the techniques that may be applied later.

When all the information has been collected and subsequently tabulated, the process of verifying hypothesis compliance for the application of the desired statistical techniques starts. Each technique, depending on their mathematical basis, requires the verification of a different set of conditions. Parametric techniques, which have higher power, are more demanding than nonparametric techniques. For this reason, sometimes the latter should be used as an alternative to the former and never be used despite their noncompliance.

After implementing these techniques to the research and obtaining the respective results, the conclusions extracted from them must be coherent so that, when inferring the results to the population, it

matches the one defined at the beginning of the research, and so that the possible existing weaknesses can be perceived and assessed. The lack of accuracy within the statements can result in misinterpretations.

FOCUS OF THE ARTICLE

The necessary information for the correct application of the techniques discussed above is detailed below. It includes a definition of each technique, as well as information regarding their utility, the objective pursued through their application, the hypothesis required and their implementation through computer software – specially R (Version 3.5.1; Ihaka & Gentleman, 2018).

Descriptive Analysis

A variable can be defined as the characteristic which is the object of a statistical study. This characteristic can be of different nature according to the values that it takes. In this way, we can make a simple classification between qualitative and quantitative variables. On the one hand, qualitative variables describe an intrinsic quality of an individual and whose value only identifies it, by characterizing it with respect to others (for instance: gender, marital status or race). On the other hand, quantitative variables offer information relative to a measurable quantity of magnitude, such as meters of a certain length or grams of a certain weight.

Once both the objectives of the statistical study, and the variables to be analysed have been defined, the next step is to obtain the information. This information may come from primary sources (that the research team obtains directly from the population under study) or from secondary sources (information available from accredited organizations that offer complete confidence to the researcher in the procedure for collecting the information as in the methodology used to process the data).

For data collection, and especially in the field of social sciences where it is frequent the use of surveys or personal interviews as research instruments, the development of the instrument and the choice of the sampling system have a central importance, since the availability of the information, that will later be analysed, depends on them. Among common problems encountered during this phase of a research we have the absence of relevant information, the confusion in the respondents' answers, the induction of bias in the answers or the low participation and disinterest of the interviewees.

It is advisable to carry out a pilot study prior to the final to evaluate elements such as the consistency shown in the responses by the interviewees; the reliability of the answers, ensuring that questions measure what they are intended to, the interviewees' degree of exhaustion according to the extension of the questionnaire, as well as the degree of complexity of the instrument, among others. All this must be evaluated simultaneously by specialists from different disciplines that converge on the topic for which the instrument has been prepared and in which the statistical study is framed.

Descriptive analysis is the study of the data that allows the description, summarization and graphic representation of the information obtained from a sample. It is usually carried out within the first phase of a statistical study and its objective is to describe the main characteristics of the individuals subject of study and the variables measured, as well as finding patterns, analysing the existence of relationship or differences between variables or groups of them and determining the existence of missing values or outliers that may have been included during the data tabulation process.

It consists of three elements that do not necessarily have to appear simultaneously in the analysis: frequency tables, graphical representations and descriptive measures. The implementation of each one

of these elements may vary depending on the nature of the analysed variable, so that the configuration and way of elaborating tables differs if the variable is qualitative or quantitative, or even if the number of modalities is high. Likewise, the use of some graphics such as a box-plot diagram or histogram makes no sense if the variable being analysed is qualitative. Neither do, as mentioned above, descriptive measures.

Frequency tables represent the distribution of a set of values according to the categories or modalities of the variable. They are used to summarize the information or to study possible relationships between two variables – this is the case of contingency tables or cross tables. In the case of qualitative or quantitative variables with a discrete and reduced set of values, they are usually developed discretely for each modality. If, on the other hand, the variable has a large number of different values, it is recommendable to combine them by creating intervals. Graphical representations have similar utilities and also allow a visual analysis of the information.

There are many descriptive measures that allow to obtain information about the data. Univariate analysis can be classified into measures of central tendency such as mean, median and mode, dispersion measures such as variance, standard deviation, range and quintiles, and shape parameters like the ones analysing the asymmetry or kurtosis of the distribution. The objective of bivariate analysis is that of determining the relationship between variables – Pearson correlation coefficient and the Spearman correlation coefficient are examples of this.

Parametric and Nonparametric Tests

Hypothesis testing is the procedure in which it is determined, based on sample statistics, whether it is acceptable (having previously set a degree of uncertainty) that a property is true for the population from which such sample came, or not. Therefore, the objective of this technique is to decide if it is possible to infer a characteristic that was noted in the sample to the whole population. In order to do this, it is necessary to establish the population previously and in an exhaustive way – also, the sample used to perform the contrast has to be randomly selected.

Parametric and nonparametric tests are based on the calculation of a sampling measure denominated test statistic, whose distribution is known so that it is possible to evaluate the probability that such statistic takes the value that has been obtained in the test if the hypothesis is true. A high probability would lead to prove the hypothesis tested. This distribution is known by the researcher, as the necessary conditions for the application of this technique guarantee it – thus ensuring the correct probability of acceptance of the hypothesis tested.

When testing a hypothesis, it is necessary to acknowledge the existence of two possibilities of error during the process: that of refuting the hypothesis tested even though it is true (previously determined probability called “significance level”) and that of accepting the hypothesis even though it is false. The probability of not committing such mistake is called “power” and it decreases when the significance level does. Two types of hypothesis test can be distinguished: parametric test and nonparametric test. The former has a higher power, which means that for given significance level, the probability of correctly accepting the hypothesis is greater. However, its implementation is more restrictive.

There are a lot of types of hypothesis test in the different fields of Statistics, most of them integrated in more complex techniques, for which their application is integrated within them. Some of the hypothesis tests most used in research articles are shown in Table 1, where it is indicated the hypothesis that is contrasted as well as the variable types for which these tests are adequate. It is necessary to emphasize that the grouping of values into categories can reduce the values of the given variable, resulting in a

lower level of measurement regarding its treatment, that is, a quantitative variable, once grouped, would be treated as qualitative ordinal, or an ordinal qualitative as qualitative nominal.

Table 1. Parametric and nonparametric hypothesis tests most used in research

Goal	Type of variable		
	Cuantitative (with normality)	Ordinal	Nominal
Normality test	Kolmogorov- Smirnov test Shapiro-Wilks test		
Compare 1 group with a theoretical value	Mean test: One sample t test	Rank test, Wilcoxon's test	Chi -Square test
Compare 2 unpaired groups	Compare means: Independent samples t test	Mann-Whitney's test	Fisher's test
Compare 2 paired groups	Compare means: Paired samples t test	Wilcoxon's test	McNemar's test
Compare k unmatched groups	Compare means: One-way ANOVA	Kruskall-Wallis test	Chi -Square test
Compare k matched groups	Compare means: Repeated measures ANOVA	Friedman's test	Cochrane's test
Quantify association between k groups	Pearson's correlation	Spearman's correlation	Contingency tables and coefficients

Principal Component Analysis and Factor Analysis

These types of analysis are framed within multivariate techniques, whose objective is to reduce the dimensionality of a data set containing a high number of variables to a smaller set of variables, uncorrelated and constructed from linear combinations. These new variables are called “components” or “factors” and they are obtained in decreasing order of importance depending on the explanatory capacity of the data. These techniques are based on the fact that, if a set of variables is strongly related, the dimensionality of such set is smaller than the number of variables on it.

The main difference between principal component analysis (PCA) and factor analysis (FA) is that, while the first analyses the total variance of the variables, FA evaluates the common variance of the variables (or covariance), so factor construction is made by maximizing the common variance and not the total variance. In addition, in PCA the components are constructed as a linear combination of original variables while the variables are explained, also as a linear combination, by factors. FA can be used to confirm which constructs underlie a set of variables - for example, in a questionnaire.

There are different procedures in order to construct components or factors. However, in all of them, minimizing dimensionality means losing explanatory capacity. Due to the organization made in the extraction process of these, such loss will be lower within the last ones, so those will be discarded. The number of components or factors selected will depend on the desired proportion of variance explained by those.

These techniques are suitable as long as the variables under study are numerical, without having to assume that they follow any concrete probability distribution. However, when the data set is distributed as a multivariate normal distribution (which is, in practice, extremely restrictive), both the analysis and the interpretation of components and factors can be done deeper. On the other hand, it is necessary that the sample size is significantly bigger than the number of variables and, obviously, a high correlation

between variables is needed. Otherwise, dimensionality reduction would be impossible and the application of these techniques would be meaningless. More information can be found in Bryant and Yarnold (1995).

An example of the application of the principal component analysis (PCA) is set forth: given a large set of socio-economic indicators of different kinds, corresponding to different countries of the world, such as the school enrolment ratio, infant mortality rate, gross domestic product (GDP), energy consumption per capita, or the annual rate of deforestation. The PCA allows us to estimate a set of variables, such as a linear combination of some of the previous ones, that can explain more complex factors and processes such as development or sustainability indexes for the different countries involved in the study, reducing that way the dimensionality of the data set.

A large part of the applications of exploratory factor analysis (FA) is carried out in the analysis of complex cognitive processes in the fields of psychology and sociology. Some characteristics such as individuals' attitudes or beliefs are concepts that cannot be directly asked in a questionnaire due to its complexity. For this reason, the researcher performs a decomposition of these characteristics into simpler components (called factors) and elaborates, for each of them, a set of items that allow them to be measured. The exploratory factor analysis allows us to test the suitability of the instrument created by the researcher for the measurement of that complex characteristic and its corresponding factors.

Causal Models

Sometimes, two types of variables can be distinguished: exogenous variables, whose values are determined by factors or variables external to the data set, and endogenous variables, which can be explained through other variables of the set. A causal model is any mathematical model that allows estimating the value of one or more endogenous variables from known values from the exogenous values. Sometimes, endogenous variables can, in turn, act as exogenous predictors of others. Therefore, the objective of this group of techniques is to estimate, with as little mistakes as possible, the values of a variable from known values of other variables, which are the cause of the first. This allow us to estimate, for instance, the probability that an individual will develop a disease based on medical tests taken or the growth of the GDP of a country before the year has finished. The ultimate aim is that of obtaining useful information that allows decision-making.

There are different types of causal models depending on the quantitative or qualitative nature of the variables. The most used ones are: regression analysis (in particular, linear regression, in which most of the variables are quantitative), logistic model, in which the dependent variable is dichotomous and in which the probability of occurrence of the event determine by it is calculated; and variance analysis, in which the endogenous variable is quantitative and the exogenous variables are nominal qualitative variables. The objective of variance analysis is that of determining how the measure of the dependent variable changes depending on the function of the different modalities of factors. Considering the simplest case, these models coincides with multiple median test.

In the case of linear regressions, the necessary assumptions for the application of these models, in addition to those considered about variable's nature, are: normality, homoscedasticity and absence of correlated errors, as well as absence of collinearity between explanatory variables. A refuted hypothesis can be rectified by the application of correction methods. The necessary conditions for the application of logistic regression are less restrictive in practice: absence of collinearity between exogenous variables, independence of errors, absence of outliers and a big enough sample. Variance analysis requires normality of endogenous variables. However, the absence of this condition can be replaced, for a factor, by applying the Kruskal-Wallis test.

Causal models are often used in all fields of science and, particularly, in the field of social sciences. Due to the main objective of these models, which is to predict the value of one or more system variables from the values taken from other variables, rather than estimating the causal model itself, non-parametric methodologies are currently proliferating focusing exclusively on the improvement of the estimation of the endogenous variable value. These methodologies use the data as the main source and are characterized by autonomous learning and by an internal function called *black box*. These techniques are included in the field of artificial intelligence.

Structural Equation Models

Structural equation models share common characteristics with factor analysis and causal models, previously analysed. Besides the variables that can be observed through data collection, we are also provided with others, called “factors” or “latent variables”, which cannot be directly measured – consequently, it has to be done through the latter as their values are absent. This technique allows to construct causal models between measurable variables and latent variables, as well as constructing causal relationship between hidden variables. In other words, this can be understood as the application of several factor analysis in which factor relationships are allowed and can also be calculated. It is usually used for the verification of causal models provided by the theory.

Therefore, the main advantage of the structural equation models (SEM) with respect to the before mentioned techniques that use the combination (PCA, FA) is that SEM allow the definition of complex systems in which different types of relationships can be proposed and estimated. These relationships are expected to be found among variables of a different nature existing in the system (both latent and observable variables). This allows to contrast theoretical relationships from empirical data. For this reason, SEM models are called confirmatory models.

The starting point in the estimation of a structural equation model is the theoretical model that the researcher wishes to contrast. From this, the researcher designs a path diagram that includes both the variables and the relationships that are part of it. From this diagram, the experimental model to be estimated can be designed, which will allow confirming the theoretical model of origin. It is at this point that the researcher must gather all the information required for the estimation of the model and, after confirming the verification of the basic assumptions required, apply the theoretical model to the data, obtaining as a result the confirmation or not of the theoretical model applied.

In order to apply a structural equation model, it is necessary to make a great number of observations, as the number of parameters that have to be estimated is, in general, very high. Moreover, a strong correlation between variables can add ad noise and damage the estimation. Multivariate normal distribution of the data is necessary for the successful implementation of this technique. This is because, in that case, it would be possible to apply a maximum likelihood method for parameters estimation. However, this hypothesis can be eased as long as the variables are symmetric and mesokurtic – in this case, other estimation methods could be applied.

STATISTICAL SOFTWARE

Descriptive Analysis

It is possible to create tables, graphics and descriptive measures by using any statistical software, spreadsheet or even by using a scientific calculator. Also, it is possible to do so by using the R software, and without the need of installing additional packages other than those by default. Commands are generally intuitive and can be consulted in any basic R manual or in James, Witten, Hastie and Tibshirani (2013).

Parametric and Non-Parametric Tests

Statistical software such as STATA or SPSS allow the execution of many of them. In most cases, the execution of a single instruction in R is sufficient in order to obtain results. Functions such as *t.test()*, *var.test()* or *prop.test()* are integrated into the basic package and allow the realization of hypothesis tests for the value of the mean, variance or proportion of a population, as well as the comparison between two populations. For more information, see James, Witten, Hastie and Tibshirani (2013).

Principal Component Analysis and Factor Analysis

There are different libraries in R that allow principal component analysis. One of them is *ade4* (Chessel, Dufour & Thioulouse, 2004). From a previously defined data table, it is only necessary to execute the function *dudi.pca()* from the library to create an object containing the analysis. Functions *inertia.dudi()*, *s.corcicle()* and *s.label()*, applied to the created object, can return, in addition, contributions of each components and two types of graphic representations, respectively.

On the other hand, in order to perform a factor analysis with R we can use the library *stats* of R Core Team (2017), which includes the function *factanal()*, with which it is possible to perform a complete factor analysis, or *loadings()*, to determine factor weight of each variable.

Causal Models

The application of causal models with R is reduced to the execution of a single command. The function *lm()* can be used to estimate linear regression models, whereas for other non-linear models, the function would be *aov()*, which includes logistic regression models, *glm()*, and variance analysis models. All of them are included in the library *stats*.

Structural Equation Models

There is specific software to calculate structural equation models such as Lisrel or AMOS, which uses the driving force of the software SPSS. We can also find libraries in R, such as *lavaan*, by Rosseel, Y. (2012), with the function *sem()*; as well as for flow chart representation, like the package *lavaanPlot* by Lishinski, A. (2018).

CONCLUSION

Data analysis is a tool that allows to obtain a large amount of information from a set of variables observed on a population, with which to optimize the context in which decision-making has to be made. However, its application must be carried out with great rigor, since the refutation of a hypothesis of the mathematical theories underlying the statistical tool could lead to wrong or pointless conclusions. Because of that it is crucial that the researcher has full control of the conditions under which the technique is applicable, as well as full knowledge of the objectives for which it was designed and the methodology that has to be followed.

In this chapter we have dealt, in a very brief and informative way, with some of the most used data analysis techniques in the field of research, such as: descriptive analysis, parametric and nonparametric hypothesis test, principal component analysis, factor analysis, causal models and structural equation models. The objective is that the reader can conceptually know them and quickly identify the necessary requirements for their application, as well as the objectives for which these techniques are applied.

FUTURE RESEARCH DIRECTIONS

The significant incidence that the refutation of the necessary hypothesis has on the results motivates the extension of this work in three directions. Firstly, it is possible to quantify, by means of a simulation, the degree of affectation that the refutation of each hypothesis has on the conclusions obtained after the application of each technique. Secondly, it would be of great interest to carry out a review of the literature in order to analyse the rate of existing research in which there is a breach of the *a priori* hypothesis, grouping them by technique or field of research. Finally, there is a great number of techniques that could not be defined in this chapter due to lack of space and that could be combined in order to complete the list presented here.

REFERENCES

- Blair, J., Czaja, R. F., & Blair, E. A. (2013). *Designing surveys: A guide to decisions and procedures*. Sage Publications.
- Bryant, F. B., & Yarnold, P. R. (1995). *Principal-components analysis and exploratory and confirmatory factor analysis Reading and understanding multivariate statistics*. American Psychological Association.
- Chessel, D., Dufour, A. B., & Thioulouse, J. (2004). The ade4 package-I- One-table methods. *R News*, 4, 5–10.
- Fienberg, S. E. (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science*, 7(2), 208–225. doi:10.1214/s/1177011360
- Fowler, F. J. Jr. (2013). *Survey research methods*. University of Massachusetts. Sage Publications.
- Holmes, T. H. (2004). Ten categories of statistical errors: A guide for research in endocrinology and metabolism. *American Journal of Physiology. Endocrinology and Metabolism*, 286(4), E495–E501. doi:10.1152/ajpendo.00484.2003 PMID:15010353

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. doi:10.1037/h0071325

Ihaka, R., & Gentleman, R. (1988). R (Version 3.5.1) [Computer software]. Retrieved from <https://www.r-project.org/>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer. doi:10.1007/978-1-4614-7138-7

Kaplan, E., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457–481. doi:10.1080/01621459.1958.10501452

Lishinski, A. (2018). *lavaanPlot: Path Diagrams for Lavaan Models via DiagrammeR*. R package version 0.5.1. <https://CRAN.R-project.org/package=lavaanPlot>

Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. doi:10.18637/jss.v048.i02

ADDITIONAL READING

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics (International Student Edition)*. Pisani, R. Purves (4th ed.). WW Norton & Company.

Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference. In *International encyclopedia of statistical science* (pp. 977–979). Springer. doi:10.1007/978-3-642-04898-2_420

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. doi:10.1007/10654-016-0149-3 PMID:27209009

Lindsey, J. K. (1996). *Parametric statistical inference*. Oxford University Press.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Allyn & Bacon/Pearson Education.

KEY TERMS AND DEFINITIONS

Causal Models: Mathematical functions that allow estimating the value or one or more variables, called “endogenous variables”, from known values of other variables that have a significant influence on the former and that are called “exogenous variables”. Its construction is done by minimizing the error made during the estimation.

Data Analysis: It is a set of techniques that allow us to examine a data set with the aim of obtaining a set of conclusions that can facilitate later decision-making.

Factor Analysis: Multivariate data analysis technique that allows to reduce the dimensionality of a high set of interrelated variables through a set of variables called “factors”, from which the original variables can be extracted by means of linear combinations of them. Their construction is done by maximizing the common variance of the variables.

Principal Component Analysis: Multivariate data analysis technique that allows to reduce the dimensionality of a high set of interrelated variables through linear combinations ordered from greater to lesser explanatory capacity of the total variance of the set, and independent of each other.

Statistical Inference: It is a set of techniques of data analysis that, together with probability calculus, allow to induce the behaviour of a population from the information provided by a sample of it.

Structural Equation Models: Confirmatory technique of data analysis that allows to demonstrate an estimate effects and causal relationships between multiple variables, hidden or observable, through a family of multivariate models.

Panel Data: A Case Study Analysis

5

Vera Costa

 <https://orcid.org/0000-0003-0801-2113>

FEUP, University of Porto, Portugal

Rui Portocarrero Sarmento

 <https://orcid.org/0000-0003-1891-5295>

Statsframe, University of Porto, Portugal

INTRODUCTION

In Statistical Data Analysis, when analyzing a dataset containing variables observed through time, Panel Data regression analysis methods are commonly used. Panel data research, also associated with longitudinal or cross-sectional time-series data ($t=1, \dots, T$), can be used in the study of varied types of entities, from companies or countries to individuals ($i=1, \dots, N$). From the perspective of data structure, spatial panel data models are the combination of conventional cross-sectional and time series data models, as represented by (Zhou & Yamaguchi, 2018):

Figure 1. Structure of panel data models

(a) Spatial panel data models			
i	t	y	x
1	1	$y_{1,1}$	$x_{1,1}$
1	2	$y_{1,2}$	$x_{1,2}$
1	3	$y_{1,3}$	$x_{1,3}$
2	1	$y_{2,1}$	$x_{2,1}$
2	2	$y_{2,2}$	$x_{2,2}$
2	3	$y_{2,3}$	$x_{2,3}$
3	1	$y_{3,1}$	$x_{3,1}$
3	2	$y_{3,2}$	$x_{3,2}$
3	3	$y_{3,3}$	$x_{3,3}$

(b) Cross-sectional data models			
i	t	y	x
1	1	$y_{1,1}$	$x_{1,1}$
2	1	$y_{2,1}$	$x_{2,1}$
3	1	$y_{3,1}$	$x_{3,1}$

(c) Time series data models			
i	t	y	x
1	1	$y_{1,1}$	$x_{1,1}$
1	2	$y_{1,2}$	$x_{1,2}$
1	3	$y_{1,3}$	$x_{1,3}$

Panel data research can provide means to control subjacent variables not observed or measured. Thus, it accounts for individual characteristics, for example, when studying the evolution of several communities or groups in social media through time, differences in behaviors across communities or variables that change over time but not across communities (i.e., global rules, agreements between communities or social media platforms rules)

DOI: 10.4018/978-1-7998-3479-3.ch045

This document is focused in two techniques used to analyze panel data:

- Fixed effects
- Random effects

Thus, the authors initiate the research by introducing the reader to the background state of the art regarding Panel Data. Then, the primary focus of the chapter considers the introduction to the case study data and obtained results. Results start with the model calibration for the Linear Regression. Additionally, the authors present and explain the case study results for four specifications, one-way or two-way, fixed or random effects, and compare the final results.

BACKGROUND

Historically, econometric and statistical models have been developed by using cross-sectional or time-series data (Washington, Karlaftis, & Mannering, 2003). However, in several cases, there is an availability of data based on cross-sections of individuals observed over time (or other observational units such as firms, geographic entities, and so on). Data which combines cross-sectional and time-series characteristics, can be called **panel data**, **pooled data** or **longitudinal data** (Dougherty, 2006).

Panel data can provide predictions on the evolution of a certain dependent variable according to other variables that are measured among distinct entities (cross-sectional) and time intervals (time series). Thus, it allows researchers to construct and test realistic behavioral models that cannot be identified using only cross-sectional or time-series data. Formally, a panel data has the following form (Kunst, 2011)

$$X_{it}, i=1, \dots, N, t=1, \dots, T.$$

Panel data can be represented in a rectangular form, like a board. Dimension i is called the “individual dimension,” and t is the time dimension. X can be a scalar (real) variable or also a vector-valued variable. Additionally, a general panel data regression model is written as (Hauser, 2013)

$$y_{it} = \beta_0 + x'_{it}\beta + \epsilon_{it}.$$

where:

x'_{it} is a K dimensional vector of explanatory variables, without a constant term,

β_0 the intercept, is independent of i and T .

β is a $(K \times 1)$ vector, the slopes, is independent of i and T .

ϵ_{it} the error, varies over i and T .

Individual characteristics (which do not vary over time), z_i may be included. In this case, the panel data regression model is written as

$$y_{it} = \beta_0 + x'_{it}\beta_1 + z'_i\beta_2 + \epsilon_{it}.$$

where z'_i is a K dimensional vector of individual characteristics (time-invariant).

Panel data analysis provide many **advantages** over conventional cross-sectional or time-series datasets (Hsiao, 2003). The main advantages are:

- First, from a statistical perspective, panel data usually give the researcher a large number of observations, which increase the degrees of freedom and reduce the collinearity among explanatory variables. Thus, in comparison with time-series data, panel data improves the efficiency of the parameter estimates.
- Second, panel data allows analyzing questions that cannot be adequately addressed using either time-series or cross-sectional data (Washington et al., 2003). In this context, the author gives the following example: “suppose that a cross-section of public transit agencies reveals that, on average, public transit subsidies are associated with 20% of increased ridership. If a homogeneous population of public transit firms is available, this might be construed to imply that a firm’s ridership will increase by 20% given transit subsidies. This might also be the implication for a sample of heterogeneous firms. However, an alternative explanation in a sample of heterogeneous public transit firms is that the subsidies have no effect (0% increase) on four-fifths of the firms, and raise ridership by 100% on one-fifth of the firms. Although these competing hypotheses cannot be tested using a cross-sectional sample (in the absence of a cross-sectional variable that “explains” this difference), it is possible to test between them by identifying the effect of subsidies on a cross-section of time series for the different transit systems.”
- Third, panel data allows researchers to check which specifications are appropriate, from the simplistic to the complex. Additionally, panel data allows accounting for cross-sectional heterogeneity by introducing additional parameters into the model (Washington et al., 2003). This way, testing for cross-sectional homogeneity is equivalent to testing the null hypothesis that these additional parameters are equal to zero.

Notwithstanding the advantages associated with panel data, when compared with cross-sectional or time-series data, panel data models have inherent disadvantages that should be reckoned. Thus, during analysis fails to account the following issues could seriously affect the estimated parameters and lead to inappropriate statistical inferences.

- First, and the most important, is **heterogeneity** bias (Hausman & Taylor, 1981). The cross-sectional units involved in the regression may have substantial differences that are not explicitly revealed by the data, leading to a model with limited or no efficiency. A strategy to account for this heterogeneity is to include additional parameters that can explicitly explain it. If heterogeneity across cross-sectional units is not considered in a statistical model, estimated parameters are biased because they capture part of the heterogeneity. Indeed, as noted by (Greene, 2012), cross-sectional heterogeneity should be the central focus of panel data analysis.
- A second issue is a **correlation in error terms**. Since an entity is repeatedly observed throughout time, it is very likely to exist a correlation between the error terms for those distinct time instants. Despite this, some disturbances associated with observations in one period are dependent on disturbances from prior time periods (Washington et al., 2003). This compromises the efficiency of Ordinary Least Squares (OLS) methods, requiring other approaches such as Generalized Least Squares (GLS).
- A third issue is the presence of heteroscedasticity, which refers to the variance of the disturbances not being constant across observations. OLS methods assume that error terms are independent and

identically distributed (iid). This might not be the case of panel data. Again, the efficiency of OLS can be compromised (Sobral & Dionísio, 2015).

In order to overcome these issues, two types of effects should be considered: **one-way** and **two-way**. One-way effects models consider one set of dummy variables (e.g., country, year), while two-way effects models consider two sets of dummy variables, in which one of them is necessarily a time dimension (e.g., country and year). Those sets are used to evaluate differences in intercept or variance of the error terms. Furthermore, in a one-way model, the error term is provided by $\epsilon_{ij} = \mu_i + v_{ij}$ where μ_i is the unobserved individual specific effect and v_{ij} is a random term. In a two-way model, the error term is given by $\epsilon_{ij} = \mu_i + \lambda t + v_{ij}$ where λt denotes the unobserved time effects in the regression (Sobral & Dionísio, 2015).

When differences are assessed in the intercept, the models are considered to be of fixed effects. However, when differences are evaluated in the variance of the error terms, the models are supposed to be of random effects. The main difference between them relies on the role of the dummy variables. Thus, if dummy variables are considered as a part of the intercept, a fixed effect model should be considered. If dummy variables act as an error term, there is considered a random effect model (Hun Myoung, 2009).

Fixed Effect Models

Fixed effect models examine group differences in intercepts, assuming the same slopes and constant variance across entities or subjects. Since a group effect is time-invariant and considered a part of the intercept, it is allowed to be correlated to other regressors (Nunes & Ferreira, 2013). Testing for the joint significance of the included fixed effects parameters (the dummy variables) is straightforwardly conducted using the Chow F test, which is given by

$$F_0 = \frac{(RRSS - URSS) / (n - 1)}{URSS / (nT - n - P)} \sim F_{n-1, n(T-1)-P}$$

for the one-way model, and is given by

$$F_0 = \frac{(RRSS - URSS) / (n + T - 2)}{URSS / ((n - 1)(T - 1) - P)} \sim F_{n+T-2, (n-1)(T-1)-P}$$

for the two-way model, where:

n is the number of cross-sectional units (individuals, counties, states, etc.),

T is the number of time periods,

P is the number of independent variables (explanatory variables).

where $RRSS$ are the restricted residual sums of squares from OLS (ordinary least squares) on the pooled model and $URSS$ are the unrestricted residual sums of squares from the model with fixed effects included (break in two points).

The null hypothesis for the test is that there is no break point (i.e., the data set can be represented by a single regression line) and, consequently, Ordinary Least Squares regression is favored. The alternative hypothesis is that fixed effects should be considered.

Random Effect Models

Random effect models estimate variance components for groups (or time) and error, assuming the same intercept and slopes. The is part of the errors and thus should not be correlated to any regressor. The difference among groups (or time periods) lies in their variance of the error term, not in their intercepts. An advantage of random effects is that it can include time-invariant variables, like age or gender. In the fixed effects model these variables are absorbed by the intercept.

Random effects are examined by Breusch and Pagan's Lagrange multiplier (LM). The null hypothesis of this test is that there are no group effects in the random effects model (Duarte, Lamounier, & Takamatsu, 2007), i.e., the pooled Ordinary Least Squares regression is favored (Nunes & Ferreira, 2013). However, if the hypothesis is rejected, it means that some effects should be considered.

Fixed Effect vs. Random Effect Models

Hausman test compares fixed effect and random effect models. Thus, computation of the Hausman statistic requires estimates of both the random and fixed effects models.

Hausman statistic has a chi-square distribution with as many degrees of freedom as there are predictors in the model (Worrall, 2010). It allows to see if there is a correlation between the individual errors and the regressors in the model, and the null hypothesis is that there is no correlation between the two (Martins, 2006). In other words, the null hypothesis of Hausman test is that the preferred model is random effects. The alternate hypothesis is that the preferred model is fixed effects.

Systematically, Table 1 presents a summary of the tests used to analyze effects of panel data models.

Table 1. Summary of statistical tests

	Test	H0	H1
Fixed effects	Chow F test	There are no groups effects (OLS regression is indicated)	There are some effects (regression with fixed effects is indicated)
Random effects	Breusch-Pagan Lagrange Multiplier test	There are no groups effects (OLS regression is indicated)	There are some effects (regression with random effects is indicated)
Fixed effects vs. Random effects	Hausman test	The preferred model is random effects	The preferred model is fixed effects

The **goodness-of-fit** of the models was assessed by the R^2 values. R^2 is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R^2 is fairly straight-forward: it is the percentage of the response variable variation that is explained by a linear model (Stone, Scibilia, Pammer, Steele, & Keller, 2018b). Thus,

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Since R^2 is a percentage, it is always between 0 and 100%, 0% indicates that the model explains none of the variability of the response data around its mean, and 100% indicates that the model explains all the variability of the response data around its mean.

However, R^2 cannot determine whether the coefficient estimates and predictions are biased because every time a predictor is added to a model, the R^2 increases, even if due to chance alone, and it never decreases. Therefore, a model with more terms may appear to have a better fit. Thus, adjusted R^2 is considered to solve this problem, since it is a modified version of R^2 that has been adjusted for the number of predictors in the model. The adjusted R^2 increases only if the new term improves the model more than would be expected by chance, and it decreases when a predictor improves the model by less than expected by chance (Stone, Scibilia, Pammer, Steele, & Keller, 2018a).

FOCUS OF THE ARTICLE

Objective

The objective of this work is to explain and exemplify how to use panel data models based on the R programming language. In this sense, the presented problem is to estimate the number of publications in a set of institutions. Therefore, the aim is to compare prediction capabilities concerning a classical regression model as benchmark. To do this, several lines of code and their output in R language are presented. Also, the most critical values obtained in the output are discussed, as well as the main conclusions are drawn.

Data

The fictitious database used in this study contains records of 246 research institutes (synthetic data), collected over six years (from 2011 to 2016). Thus a total of 1476 instances are available. For each research institution in each year, several variables are considered such as:

- Researchers - represents the average number of researchers
- Supervisors - represents the average number of supervisors
- Publications - represents the total number of publications
- Rcenters - represents the number of research center
- Projects - represents the total number of projects

R software was used to explore the presented analysis. So, the first ten rows of the dataset could be obtained by the expression below.

```
Computer Code      head(Panel, 10)
Output   Institution Year Researchers Supervisors Publications Rcenters Proj-
ects
1         1 2011      12.2         6.0          25          3         3
2         1 2012      15.0         6.0          30          3         3
3         1 2013      14.0         5.0          37          3         5
4         1 2014      14.2         6.0          37          3         5
5         1 2015      14.7         6.0          39          3         5
```

6	1	2016	16.4	5.2	44	3	5
7	2	2011	17.2	7.0	39	3	5
8	2	2012	17.2	8.4	43	3	5
9	2	2013	18.9	8.7	49	3	6
10	2	2014	20.2	8.7	54	3	7

In order to understand the data variation, some simple statistics should be performed, namely, the counting of discrete variables – Institution and Year – and a descriptive analysis of continuous variables (mean, maximum and minimum, median and quartiles) – Researchers, Supervisors, Publications, Rcenters, and Projects.

```
Computer Code      ##counting of Institutions
length(unique(Panel$Institution))

##counting of Years
length(unique(Panel$Year))

## descriptive analysis of Researchers, Supervisors, Publications, Rcenters
and Projects variables
summary(Panel[,3:7])
Outputs           [1] 246

[1] 6

  Researchers    Supervisors    Publications    Rcenters    Projects
Min.   : 1.40   Min.   : 1.70   Min.   : 6.00   Min.   :2.000   Min.   : 1.000
1st Qu.:11.10   1st Qu.: 6.50   1st Qu.:23.00   1st Qu.:2.000   1st Qu.:
2.000
Median:16.90   Median:10.00   Median: 37.00   Median:3.000   Median: 4.000
Mean   :19.08   Mean   :11.03   Mean   : 43.84   Mean   :2.715   Mean   : 5.383
3rd Qu.:23.90   3rd Qu.:14.00   3rd Qu.: 54.25   3rd Qu.:3.000   3rd Qu.:
7.000
Max.    :79.20   Max.    :43.10   Max.    :242.00   Max.    :5.000   Max.    :37.000
```

Based on the outputs, it is possible to observe that 246 institutions and six years are effectively presented in the database. Regarding continuous variables, a high variation is observed in most variables. Rcenters is the variable with less variation since it varies from 2 to 5.

As the objective of this study is to find a model to estimate the number of publications, Publications variable will be considered a dependent variable. This variable varies from 6 to 242 with a mean equal to 43.84 and a median equal to 37, which may indicate that the research center with 242 publications could be an outlier of the number of publications.

Researchers, Supervisors, Rcenters, and Projects will be used in the model as independent variables.

Methodology

The present study has a main objective to analyze some models of panel data and verify which is the best to estimate the number of publications. In order to do this, several models were tested using R software. The R code used to make all analysis is only an example. Probably, other packages or other expressions could be used. However, in the present work, in order to estimate panel data regression models, plm package in R was used (Croissant & Millo, 2008).

The first phase consists of the discussion of the influence of the independent variables (IV). To get conclusions about this, the input information was the correlation matrix, and the significance of the coefficients results from the tested models. After narrowing the IV that is significant to the model, the chosen model was tested with the selected variables in order to test for improvements.

Thus, Table 2 presents a summary of the panel data models used in this work, as well as a description of each model to a best understanding. In addition to that model collection, an initial OLS model was built. For instance, if the statistical analysis indicates that there are no relevant fixed or random effects, it is likely that a regular OLS model will suffice for the estimation.

Table 2. Panel data models

Model		Description
1	OLS	Considers only the variables, create a regular linear regression with no group effects
2	Fixed effects – one-way error (by institution)	Considers that there are fixed effects only in one dimension and creates the model by grouping by institution
3	Fixed effects – one-way error (by year)	As model 2, considers fixed effects in one dimension, but groups by year.
4	Fixed effects – two-way error	Add both dimensions, grouping by farm and year.
5	Random effects – one-way error (by institution)	The same as model 2, but resorts to random effects.
6	Random effects – one-way error (by year)	The same grouping of model 3, using random effects.
7	Random effects – two-way error	The same as model 4, also considering random effects.

Results

As mentioned before, in order to analyze the best model to estimate the number of publications, it is important to verify how variables are correlated between them. Thus, the R code and the output of the correlation matrix of the variables in study is presented. The results improve the understanding of how the IV are related to the DV (dependent variable) – Publications – and if the IV are related to each other.

```
Computer Code      library("Hmisc")
```

```
correlation_matrix <- rcorr(as.matrix(Panel[,3:7]))
```

```
correlation_matrix
```

```
Output      Researchers Supervisors Publications Rcenters Projects
Researchers 1.00        0.70        0.95        0.52        0.90
Supervisors 0.70        1.00        0.62        0.32        0.56
Publications 0.95        0.62        1.00        0.50        0.95
```

Rcenters	0.52	0.32	0.50	1.00	0.46
Projects	0.90	0.56	0.95	0.46	1.00

n= 1476

P

	Researchers	Supervisors	Publications	Rcenters	Projects
Researchers	-	0	0	0	0
Supervisors	0	-	0	0	0
Publications	0	0	-	0	0
Rcenters	0	0	0	-	0
Projects	0	0	0	0	-

From that analysis, it is visible that Researcher and Projects have the most substantial influence in the DV ($corr=0.95$, $p<0.0001$). It was also noticeable that Rcenters has a lower correlation with DV ($corr=0.50$, $p<0.0001$). However, the main conclusion obtained from the presented output is that all independent variables are significantly correlated to the dependent variable, which may indicate a good choice of independent variables to define the panel data model regression.

Model 1. OLS

```
Computer Code      library(plm)
inst.fixed.lm = lm(Publications ~ Researchers + Supervisors + Rcenters + Proj-
ects, data = Panel)
summary(inst.fixed.lm)
Output Call:
lm(formula = Publications ~ Researchers + Supervisors + Rcenters +Projects,
data = Panel)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-33.965	-3.294	0.135	3.301	29.982

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.19433	0.85438	-0.227	0.8201
Researchers	1.25933	0.04319	29.155	<2e-16 ***
Supervisors	-0.03108	0.04072	-0.763	0.4455
Rcenters	0.57102	0.34309	1.664	0.0963 .
Projects	3.49239	0.08453	41.315	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.68 on 1471 degrees of freedom

Multiple R-squared: 0.9532, Adjusted R-squared: 0.9531

F-statistic: 7488 on 4 and 1471 DF, p-value: < 2.2e-16

Results of the OLS model shows the variation of residuals, the coefficients, and some important statistics. First, the $Pr(>|t|)$ is a two-tail p-value that test the hypothesis that each coefficient is equal to 0. Since Researchers and Projects variables presents coefficients with a $Pr(>|t|) < 0.05$, the null hypothesis is rejected. Thus, at a significance level of 5%, Researchers and Projects have coefficients different of zero and contributes to the model (eventually, Rcenters could be considered at a significance level of 10%). Additionally, the difference in the coefficients of Researcher and Projects could indicate that Projects have a higher influence on the DV (higher coefficient value). The value of t also is a good indicator, since the higher the t-value the higher the relevance of the variable.

Second, results of R2 and adjusted R2 are 0.9532 and 0.9531, respectively, which means that approximately 95% of the total variability of Publications is explained by the predictors.

Finally, the output shows the result of a F-test, which test if all coefficients are equal to zero (Torres-reyna, 2010). The model presents an F value of 7488 with 4 and 1471 degrees of freedom, with $p < 0.0001$. The result indicate the null hypothesis should be rejected at a 5% of significance level, indicating that some coefficient is different from zero and, consequently, the model is well defined.

Model 2. Fixed effects – one-way error (by institution)

```
Computer Code      library(plm)
inst.fixed = plm(Publications ~ Researchers + Supervisors + Rcenters + Proj-
ects, data = Panel, index = c("Institution"), model = "within")
summary(inst.fixed)
Output Oneway (individual) effect Within Model
```

Call:

```
plm(formula = Publications ~ Researchers + Supervisors + Rcenters +
    Projects, data = Panel, model = "within", index = c("Institution"))
```

Balanced Panel: n=246, T=6, N=1476

Residuals:

	Min.	1st Qu.	Median	3rd Qu.	Max.
	-27.0000	-1.7500	0.0291	1.8000	21.5000

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
Researchers	1.429442	0.059285	24.1112	<2e-16 ***
Supervisors	0.067189	0.057303	1.1725	0.2412
Rcenters	0.449300	0.456434	0.9844	0.3251
Projects	3.067028	0.089962	34.0924	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 168010

Residual Sum of Squares: 21945

R-Squared: 0.86938

Adj. R-Squared: 0.72213

F-statistic: 2040 on 4 and 1226 DF, p-value: < 2.22e-16

The second model (fixed effects, one-way error by institution) shows that only Researchers and Projects have a significant coefficient value. Similar to the first model, Projects have the highest coefficient value and t-value, which could indicate that Projects have a higher influence on the number of Publications (the dependent variable).

Values of R² and adjusted R² are 0.86938 and 0.72213, which indicates that approximately 72% of the total variability of Publications is explained by the predictors.

The model presents an F value equal to 2040 with 4 and 1226 degrees of freedom. Since the p-value is lower than 0.0001, the null hypothesis that all coefficients are equal to zero is rejected, and the model is well defined.

Comparing with the first model, this model seems to be worst, because of the values of R² and adjusted R². However, some specific tests will be performed to indicate the most suitable model.

Model 3. Fixed effects – one-way error (by year)

```
Computer Code      year.fixed = plm(Publications ~ Researchers + Supervisors +
Rcenters + Projects, data = Panel, index = c("Year"), model = "within")
summary(year.fixed)
Output Oneway (individual) effect Within Model
```

Call:

```
plm(formula = Publications ~ Researchers + Supervisors + Rcenters +
      Projects, data = Panel, model = "within", index = c("Year"))
```

Balanced Panel: n = 6, T = 246, N = 1476

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-33.04541	-3.31680	0.15421	3.30897	28.95882

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
Researchers	1.273612	0.042958	29.6480	< 2e-16 ***
Supervisors	-0.026765	0.040426	-0.6621	0.50804
Rcenters	0.706666	0.341397	2.0699	0.03863 *
Projects	3.405167	0.085507	39.8231	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 1328600

Residual Sum of Squares: 64394

R-Squared: 0.95153

Adj. R-Squared: 0.95124

F-statistic: 7195.36 on 4 and 1466 DF, p-value: < 2.22e-16

Panel Data

The third model presents R^2 and adjusted R^2 very similar, and approximately 95%. Regarding the coefficients of the independent variables, Researchers, Rcenters and Projects have $p < 0.05$, i.e., they contribute significantly to the model. Similar to the previous models, F-test also presents a $p < 0.001 < 0.05$, indicating the model is well defined.

Model 4. Fixed effects – two-way error

```
Computer Code      two.fixed = plm(Publications ~ Researchers + Supervisors +
Rcenters + Projects, data = Panel, index = c("Institution", "Year"), effect =
"twoways", model = "within")
summary(two.fixed)
OutputTwoways effects Within Model
```

Call:

```
plm(formula = Publications ~ Researchers + Supervisors + Rcenters +
    Projects, data = Panel, effect = "twoways", model = "within",
    index = c("Institution", "Year"))
```

Balanced Panel: n = 246, T = 6, N = 1476

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-24.669304	-1.618241	0.075276	1.678866	20.595028

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
Researchers	1.230850	0.058511	21.0363	<2e-16 ***
Supervisors	0.066872	0.054335	1.2307	0.2187
Rcenters	0.295760	0.432503	0.6838	0.4942
Projects	2.853220	0.086980	32.8033	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 94551

Residual Sum of Squares: 19564

R-Squared: 0.79309

Adj. R-Squared: 0.75004

F-statistic: 1170 on 4 and 1221 DF, p-value: < 2.22e-16

The fourth model analyses the two-way error with fixed effects and presents an R^2 and adjusted R^2 of 0.79309 and 0.75004 respectively. Coefficients of variables Researchers and Projects are those that contribute significantly to the model. Additionally, the F-test indicates that, similarly to the other models, this one is also well defined.

Model 5. Random effects – one-way error (by institution)

```
Computer Code      inst.random = plm(Publications ~ Researchers + Supervisors +
Rcenters + Projects, data = Panel, index = c("Institution"), model = "random")
summary(inst.random)
```

```
OutputOneway (individual) effect Random Effect Model
(Swamy-Arora's transformation)
```

Call:

```
plm(formula = Publications ~ Researchers + Supervisors + Rcenters +
Projects, data = Panel, model = "random", index = c("Institution"))
```

Balanced Panel: n = 246, T = 6, N = 1476

Effects:

```
              var std.dev share
idiosyncratic 17.900    4.231 0.401
individual    26.749    5.172 0.599
theta: 0.6832
```

Residuals:

```
      Min.      1st Qu.      Median      3rd Qu.      Max.
-29.357049 -1.815641   0.052529   1.836900   23.440526
```

Coefficients:

```
              Estimate Std. Error t-value Pr(>|t|)
(Intercept) -0.919175   1.111322 -0.8271   0.4083
Researchers   1.369133   0.046773 29.2721 <2e-16 ***
Supervisors   0.015672   0.047834  0.3276   0.7432
Rcenters      0.484708   0.387836  1.2498   0.2116
Projects      3.185626   0.081094 39.2829 <2e-16 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Total Sum of Squares: 291830

Residual Sum of Squares: 26450

R-Squared: 0.90936

Adj. R-Squared: 0.90912

F-statistic: 3689.65 on 4 and 1471 DF, p-value: < 2.22e-16

Similar to the previous models, the random effects model (one-way error by institution) output gives some descriptive statistics and the result of some statistical tests. The main results to take into account are the p-value of the coefficients, which indicates variables that contribute to the model. In this case, only Researchers and Projects should be considered. R2 and adjusted R2 indicates a good value of goodness-of-fit, and the p-value of the F-test indicate the model is well defined.

Panel Data

Model 6. Random effects – one-way error (by year)

```
Computer Code      year.random = plm(Publications ~ Researchers + Supervisors +
Rcenters + Projects, data = Panel, index = c("Year"), model = "random")
summary(year.random)
```

```
OutputOneway (individual) effect Random Effect Model
(Swamy-Arora's transformation)
```

Call:

```
plm(formula = Publications ~ Researchers + Supervisors + Rcenters +
Projects, data = Panel, model = "random", index = c("Year"))
```

Balanced Panel: n = 6, T = 246, N = 1476

Effects:

	var	std.dev	share
idiosyncratic	43.925	6.628	1
individual	0.000	0.000	0
theta:	0		

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-33.96527	-3.29422	0.13542	3.30080	29.98220

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	-0.194328	0.854379	-0.2275	0.82011
Researchers	1.259329	0.043195	29.1547	< 2e-16 ***
Supervisors	-0.031075	0.040717	-0.7632	0.44546
Rcenters	0.571023	0.343094	1.6643	0.09626 .
Projects	3.492393	0.084530	41.3155	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 1402100

Residual Sum of Squares: 65632

R-Squared: 0.95319

Adj. R-Squared: 0.95306

F-statistic: 7488.37 on 4 and 1471 DF, p-value: < 2.22e-16

The sixth model presents an R² and adjusted R² of 0.95319 and 0.95319 indicating that approximately 95% of the total variability of Publications is explained by the predictors.

Once again, the model is well defined (p-value of the F-test lower than 0.0001 < 0.05). Coefficients of variables Researchers and projects contribute significantly to the model. In particular, Rcenters variable could also be included in the model if the level of significance considered was 10%.

Model 7. Random effects – two-way error

```
Computer Code      two.random = plm(Publications ~ Researchers + Supervisors +
Rcenters + Projects, data = Panel, index = c("Institution", "Year"), effect =
"twoways", model = "random")
summary(two.random)
OutputTwoways effects Random Effect Model
(Swamy-Arora's transformation)
```

Call:

```
plm(formula = Publications ~ Researchers + Supervisors + Rcenters +
Projects, data = Panel, effect = "twoways", model = "random",
index = c("Institution", "Year"))
```

Balanced Panel: n = 246, T = 6, N = 1476

Effects:

```
              var std.dev share
idiosyncratic 16.023   4.003 0.372
individual    27.062   5.202 0.628
time          0.000   0.000 0.000
theta: 0.7003 (id) 0 (time) -5.551e-17 (total)
```

Residuals:

```
      Min.      1st Qu.      Median      3rd Qu.      Max.
-29.222046 -1.841300   0.050933   1.831155  23.289466
```

Coefficients:

```
              Estimate Std. Error t-value Pr(>|t|)
(Intercept) -0.960335   1.129840 -0.8500   0.3955
Researchers   1.372851   0.047099 29.1485 <2e-16 ***
Supervisors   0.018847   0.048149  0.3914   0.6955
Rcenters      0.479157   0.389774  1.2293   0.2191
Projects      3.176388   0.081070 39.1807 <2e-16 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Total Sum of Squares: 278850

Residual Sum of Squares: 25983

R-Squared: 0.90682

Adj. R-Squared: 0.90657

F-statistic: 3578.94 on 4 and 1471 DF, p-value: < 2.22e-16

The last model presents an R² and adjusted R² of 0.90682 and 0.90657 respectively. The model is well defined since the p-value of the F-test is lower than 0.05, and the variables that should be considered in the model are Researchers and Projects.

Statistical Tests

In order to test the best models, some statistical test may be performed. As described in the previous section, three tests need to be used:

Chow F-test - to evaluate the significance of fixed effects (alternative hypothesis)

```
Computer Code      pFtest(inst.fixed, inst.fixed.lm)
pFtest(year.fixed, inst.fixed.lm)
pFtest(two.fixed, inst.fixed.lm)
Output            F test for individual effects

data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
F = 9.9618, df1 = 245, df2 = 1226, p-value < 2.2e-16
alternative hypothesis: significant effects
```

F test for individual effects

```
data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
F = 5.6365, df1 = 5, df2 = 1466, p-value = 3.734e-05
alternative hypothesis: significant effects
```

F test for twoways effects

```
data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
F = 11.5, df1 = 250, df2 = 1221, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Previous output shows that all p-values of the F-test are lower than 0.05, which leads to conclude that the null hypothesis that there are no break point, i.e., the OLS model is the best, is rejected. The alternative hypothesis is that fixed effects should be considered. This is the conclusion of these tests.

Breusch-Pagan Lagrange Multiplier test - to evaluate the significance of random effects (alternative hypothesis)

```

Computer Code      plmtest(inst.random, type=c("bp"))
plmtest(year.random, type=c("bp"))
plmtest(two.random, type=c("bp"))
Output            Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels

data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
chisq = 1286, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects

```

```

Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels

data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
chisq = 33.048, df = 1, p-value = 8.992e-09
alternative hypothesis: significant effects

```

```

Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels

data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
chisq = 1286, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects

```

Regarding the Breusch-Pagan Lagrange Multiplier test, all of them gives p-value < 0.05 which leads to rejecting the null hypothesis, i.e., the preference for the OLS. Thus, these tests allow deciding to consider some effects. Now, it is important to decide between fixed effects or random effects.

- **Hausman test** - in case the null hypothesis is rejected in both previous tests, the test will point the recommended direction towards the best model specification, which is used to assess whether there is a correlation between the unobserved effects and the regressors. If there is, a fixed effects model will be preferred.

```

Computer Code      phtest(inst.fixed, inst.random)
phtest(year.fixed, year.random)
phtest(two.fixed, two.random)
Output            Hausman Test

data:  Publications ~ Researchers + Supervisors + Rcenters + Projects
chisq = 10.935, df = 4, p-value = 0.02731
alternative hypothesis: one model is inconsistent

```


Panel Data

Hausman Test

```
data: Publications ~ Researchers + Supervisors + Rcenters + Projects
chisq = 34.936, df = 4, p-value = 4.788e-07
alternative hypothesis: one model is inconsistent
```

Hausman Test

```
data: Publications ~ Researchers + Supervisors + Rcenters + Projects
chisq = 178.96, df = 4, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

All tests give a p-value lower than 0.05. Similar to previous tests, the null hypothesis is rejected. Thus, the preference for fixed effects instead of random effects should be considered.

Summary of the Models

Table 3 presents a summary of the main values obtained in the analysis performed in the present study. Based on this table, some conclusions should be defined.

Table 3. Summary of panel data models tested

	OLS	One-way (institution)	One-way (Year)	Two-way	One-way (institution)	One-way (Year)	Two-way
Fixed effects							
Constant	-0.19433	-	-	-	-0.919175	-0.194328	-0.960335
Researchers	1.25933***	1.429442***	1.273612***	1.230850***	1.369133***	1.259329***	1.372851***
Supervisors	-0.03108	0.067189	-0.026765	0.066872	0.015672	-0.031075	0.018847
Rcenters	0.57102	0.449300	0.706666*	0.295760	0.484708	0.571023	0.479157
Projects	3.49239***	3.067028***	3.405167***	2.853220***	3.185626***	3.492393***	3.176388***
Statistics							
R²	0.9532	0.86938	0.95153	0.79309	0.90936	0.95319	0.90682
Adjusted R²	0.9531	0.72213	0.95124	0.75004	0.90912	0.95306	0.90657
F (p-value)	-	9.9618 (<0.0001)	5.6365 (<0.0001)	11.5 (<0.0001)	-	-	-
LM (p-value)	-	-	-	-	1286 (<0.0001)	33.048 (<0.0001)	1286 (<0.0001)
H (p-value)	-	-	-	-	10.935 (0.02731)	34.936 (<0.0001)	178.96 (<0.0001)

As described before, fixed effects should be considered as the best option to the estimate publications. Looking at R^2 and adjusted R^2 values it is possible to decide that one-way error (by year) with fixed effects should be considered as the best model to estimate the number of publications.

SOLUTIONS AND RECOMMENDATIONS

As a recommendation in dealing with the consequent lack of knowledge of R language for statistical data analysis, the reader might want to read the throughout R language application to statistics in (Sarmiento & Costa, 2017).

FUTURE RESEARCH DIRECTIONS

As future research direction, it can be useful for the researcher to explore other R packages available in CRAN repository, with the purpose to explore and analyze Panel Data.

CONCLUSION

Following the state of the art introduction, a throughout research template for Panel Data is written in this chapter. With the proposal of an example case study, the reader will hopefully appreciate the advantages within the use of Panel Data to explore his/her datasets with temporal characteristics.

The authors presented a step by step approach to Panel Data, giving extreme importance to tasks like testing fixed or random effects, either individually for each variable or by using a two-way approach.

ACKNOWLEDGMENT

Vera Costa gratefully acknowledges funding from FCT (Portuguese Foundation for Science and Technology) through a Ph.D. grant (Ref. PD/BD/128065/2016). Rui Portocarrero Sarmiento also gratefully acknowledges funding from FCT (Portuguese Foundation for Science and Technology) through a Ph.D. grant (SFRH/BD/119108/2016). The authors want to thank also to the reviewers, for the constructive reviews provided in the development of this publication.

REFERENCES

- Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, 27(2), 1–43. doi:10.18637/jss.v027.i02
- Dougherty, C. (2006). Introduction to Panel Data Models. In *Introduction to Econometrics* (pp. 408–423). Academic Press.

- Duarte, P. C., Lamounier, W. M., & Takamatsu, R. T. (2007). Modelos Econométricos para Dados em Panel: Aspectos Teóricos e Exemplos de Aplicação à Pesquisa em Contabilidade e Finanças. *7º Congresso USP de Controladoria e Contabilidade*. doi:10.1007/978-3-319-31960-5
- Greene, W. H. (2012). *Econometric Analysis* (Vol. 3). Prentice Hall.
- Hauser, M. (2013). *Panel Data Models*. Spatial Econometrics. doi:10.1016/b978-0-12-813387-3.00015-9
- Hausman, J. A., & Taylor, W. E. (1981). Panel Data and Unobservable Individual Effects. *The Econometric Society*, 49(6), 1377–1398. doi:10.2307/1911406
- Hsiao, C. (2003). Introduction to panel data. In *Analysis of panel data - Second Edition* (pp. 1–11). doi:10.1016/S0898-1221(04)90099-5
- Hun Myoung, P. (2009). *Linear Regression Models for Panel Data Using SAS, Stata, LIMDEP, and SPSS*. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.
- Kunst, R. M. (2011). *Econometric Methods for Panel Data - Part I. Econometric Analysis of Panel Data*. Retrieved from <http://homepage.univie.ac.at/robert.kunst/panpres.pdf>
- Martins, L. F. (2006). *Modelos com dados de painel*. Iscte - Eg.
- Nunes, A. A., & Ferreira, M. C. (2013). *Panel Data Models and Spatial Regression Models*. Academic Press.
- Sarmiento, R., & Costa, V. (2017). *Comparative Approaches to Using R and Python for Statistical Data Analysis*. IGI Global. doi:10.4018/978-1-68318-016-6
- Sobral, T., & Dionísio, T. (2015). *A Panel Data model for Estimating Gasoline Demand*. Academic Press.
- Stone, B. K., Scibilia, B., Pammer, C., Steele, C., & Keller, D. (2018a). *Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables*. Academic Press.
- Stone, B. K., Scibilia, B., Pammer, C., Steele, C., & Keller, D. (2018b). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* Academic Press.
- Torres-Reyna, O. (2010). Getting Started in Fixed/Random Effects Models using R. *Data & Statistical Services*, 28. <http://dss.princeton.edu/training/P>
- Washington, S. P., Karlaftis, M. G., & Mannering, F. L. (2003). *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press LLC., doi:10.1201/9780203497111
- Worrall, J. L. (2010). A user-friendly introduction to panel data modeling. *Journal of Criminal Justice Education*, 21(2), 182–196. doi:10.1080/10511251003693702
- Zhou, X., & Yamaguchi, Y. (2018). Relative importance of climatic and anthropogenic drivers on the dynamics of aboveground biomass across agro-ecological zones on the Mongolian Plateau. *Sustainability (Switzerland)*, 10(3235). Advance online publication. doi:10.3390/u10103435

ADDITIONAL READING

- Baltagi, B. (2005). *Econometric Analysis of Panel Data* (3rd ed.). John Wiley & Sons, Ltd.
- Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, 3(1), 133–153. doi:10.1017/psrm.2014.7
- Frees, E. (2004). *Longitudinal and Panel Data, Analysis and Applications in the Social Sciences*. Cambridge University Press. doi:10.1017/CBO9780511790928
- Hsiao, C. (2003). *Analysis of Panel Data* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511754203
- Mátyás, L., & Sevestre, P. (1996). *The Econometrics of Panel Data: A Handbook of the Theory with Applications* (Vol. 33). Advanced Studies in Theoretical and Applied Econometrics. doi:10.1007/978-94-009-0137-7_4
- Wooldridge, J. (2001). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

KEY TERMS AND DEFINITIONS

Fixed Effects: Variables that are constant across individuals, i.e., do not change, or change at a constant rate over time (for example, age, sex, or ethnicity).

OLS Regression: Commonly called linear regression. OLS method corresponds to minimizing the sum of square differences between the observed and predicted values.

One-Way Error Model: A model that considers only one group of the data (for example, year or country).

Panel Data: Data derived from a small number of observations over time on a large number of cross-sectional units (for Example, individuals, households, firms, or governments).

Random Effects: Variables that are random and unpredictable (for example, cost of a new car varies when purchased in different years).

Two-Way Error Model: A model that considers two group of the data (for example, year and country).

Applications of Decision Support Systems in Aviation

Tetiana Shmelova

 <https://orcid.org/0000-0002-9737-6906>

National Aviation University, Ukraine

Yuliya Sikirda

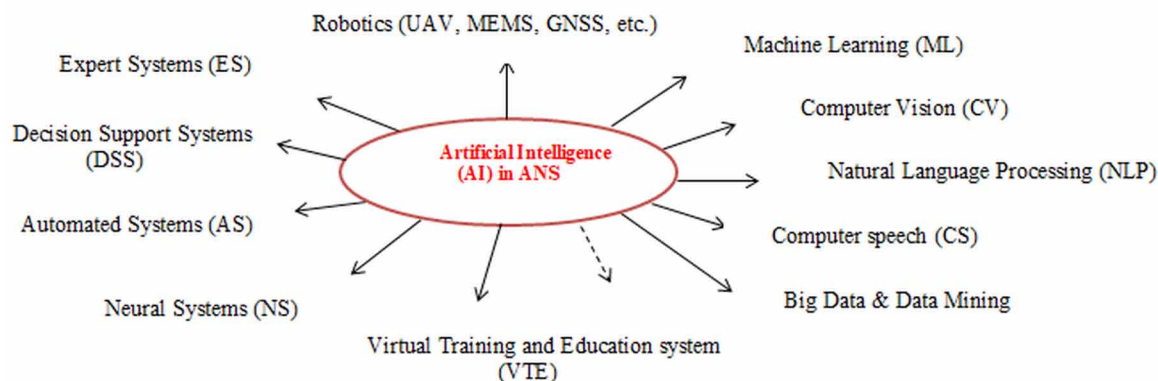
 <https://orcid.org/0000-0002-7303-0441>

Flight Academy, National Aviation University, Ukraine

INTRODUCTION

The air transport industry plays a major role in world economic activity and to maintain the safe and efficient operation of aviation enterprises that allows maximum use to be made of enhanced capabilities provided by technical advances. Nowadays the International Civil Aviation Organization (ICAO) has added in its documents new approaches for achieving the main goals of the organization, which are: to enhancing the effectiveness of global aviation security and to improve the practical and sustainable implementation of preventive aviation security measures. These new approaches include the development of progressive technology and human capability, the improvement of technological resources to improve the quality of decision making of aviation personal (pilots of manned and unmanned aircraft, air traffic controllers, engineers, etc. (ICAO, 2017; ICAO, 2018). The quality of decision dependences is using innovative technology in aviation such as Artificial Intelligence (AI). The AI document “White paper” of IATA (International Air Transport Association) presents the results of IATA research and development activities on AI in collaboration with airlines. The new technologies of AI can be clustered in the following capabilities, such as Machine learning (ML), Natural Language Processing (NLP), Expert Systems, Vision, Speech, Planning, Robotics (IATA, 2016; ICAO, 2018). To increase the productivity of work at each stage of the life cycle (LC) of aviation systems, different AI technologies and methods can be effectively applied (Figure 1).

Figure 1. The synergetic effect - LC of aviation technique with using AI capability



DOI: 10.4018/978-1-7998-3479-3.ch046

Artificial Intelligence is the simulation of human intelligence processes by modeling, computer systems, and machines. These processes include learning (the acquisition of information and rules for using the information); reasoning, estimation, and modeling (using rules to reach conclusions (approximate or definite results)); self-correction (estimation of obtained models). Particular applications of AI include Expert System (ES); Decision Support System (DSS); Automated systems; systems of pattern recognition, speech recognition, and machine vision, etc. (ICAO, 2017; IATA, 2016). AI (and DSS) are effective for minimization risks and improve the quality of decisions in the aviation system.

BACKGROUND

Support for the safe functioning of ANS is one of the most important scientific and technical problems. Statistical data show that human errors account for up to 80% of all aviation accidents (ICAO, 2004). According to ICAO data (ICAO, 2013), compared with 2011, in 2012 the number of incidents in the world decreased by 21%, in 2013 – on 13%; the number of fatalities in 2012 decreased by 10%, in 2013 – on 55%. Consequently, the least number of deaths has been fixed in 2012-2013, starting from 2004. As a result of the decreased number of incidents, an increased number of departures, the frequency of incidents in the world in 2012 has been reduced to 3,2 events per million departures and in 2013 – to 2,8 events per million departures. This is the lowest value since the ICAO started to monitor the frequency of incidents in the world. Aviation systems cannot be wholly free from dangerous factors and connected with them risks, while, the elimination of aviation events and serious incidents continues to be the final goal of human activity in the sphere of aviation safety. Neither human activity nor systems created by it guarantee a total absence of operating errors and their consequences (ICAO, 2013). In such a way, safety is a dynamic characteristic of aviation with the help of which risk factors for flight safety should steadily decrease. It is important to note that the adoption of efficiency indices of ensuring flight safety is frequently influenced by internal and international standards and also by cultural features (ICAO, 2009). While risk factors for flight safety and operating errors are under control, such an opened and dynamic system as civil aviation may be controlled providing the necessary balance between flight performance and safety requirements.

Latest demands of international aviation organizations directed on the implementation of an integrated approach for the improvement of aviation safety. Many ICAO documents describe the problems in aviation and include recommend the creation of support for a solution in each professional situation (ICAO, 2005a; 2005b; 2009). One of the ways to increase safety by timely is the support of pilots in emergency situations (ICAO, 2012). A modern approach, founded on the characteristics (Performance-Based Approach – PBA) (ICAO, 2009), based on the next three principles:

- a strong focus on desired/required results;
- informed decision making (DM) driven by those desired/required results;
- reliance on facts and data for DM.

Herein the principle “using facts and data while DM” admits that tasks shall comply with the widely known in Western management criteria **SMART**, which correspond to the abbreviation of five English words: Specific, Measurable, Achievable, Relevant, Time-bound.

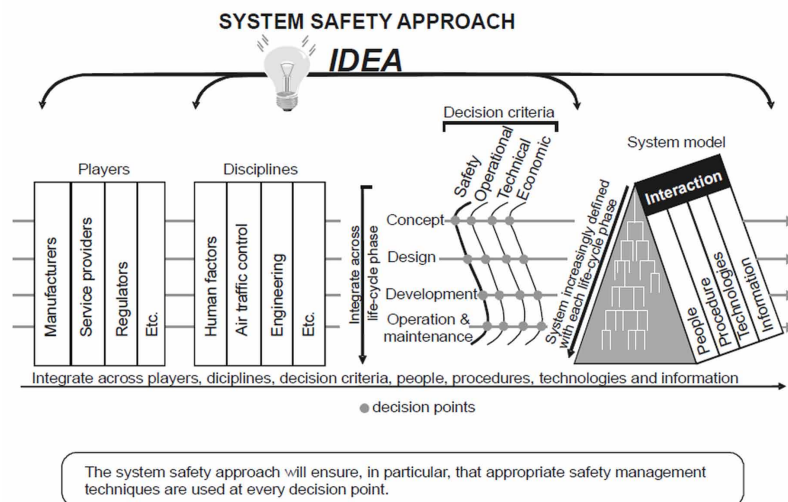
So-called aeronautical DM is DM in a unique environment – aviation. It is a systematic approach to the mental process used by pilots to consistently determine the best course of action in response to a

given set of circumstances. It is what a pilot intends to do, on the basis of the latest information he has (Federal Aviation Administration [FAA], 2008). So, ICAO constantly develops and improves proactive, based on the risks evaluation, methods, directed on the farther decrease in a number of aviation events in the world. Also, ICAO encourages aviation communities to recognize the importance of adherence to the single global approach for safety improvement and monitoring (ICAO, 2013). Manual of ICAO “Global Air Traffic Management Operational Concept” (ICAO, 2005b) presents basic methods and benefits of the support of aviation operators in their professional operations for using DSS for difficult or emergency situations. Documents of ICAO have presented concepts about the Flight & Flow Information for a Collaborative Environment (FF-ICE), including support for a Performance-Based Approach (PBA), Collaborative Decision Making (CDM), System-Wide Information Sharing and Management (SWIM) by trajectory (ICAO, 2017; ICAO, 2018). To ensure effective CDM, it is necessary to create a DSS’s for the operators of the aviation system as a whole.

Expected benefits of this approach such as: from an *airspace user* (pilots), safety systems will be perspective, greater equity in airspace access, greater access to timely and meaningful information for decision support, and more autonomy in DM, including conflict management, will provide the opportunity to better deliver business and individual outcomes within an appropriate safety framework; from a *service provider* (ATM), safety systems will be perspective, including that of aerodrome operators, the ability to operate within an information-rich environment, with real-time data, as well as system trend and predictive data, fused with a range of automated decision-support or DM tools, will enable optimization of services to airspace users; from a *regulatory perspective*, safety systems will be robust and open, allowing safety not only to be more easily measured and monitored, but also compared and integrated on a global basis, not for its own sake, but as a platform for continuous improvement.

Maximum effectiveness and efficiency of actions can be obtained by activities undertaken in the early phases of any system’s life cycle since correcting problems during requirements definition and design is generally the most effective. Figure 2 depicts this process for the ATM system as a whole. It describes how, starting from the conceptual level and even during operation, different players and disciplines make decisions against different criteria – including *safety*. These decisions will progressively define the ATM system (i.e. the way in which it is organized and the way in which people, procedures, technologies and information interact to perform a task) for using different DSS on all stages of operating ATM system.

Figure 2. System safety approach (ICAO, 2009)



A lot of research is being done for improving the work of DSSs. The problems of DSS's informational provision have been considered in the works of scientists (Tarasov, & Gerasimov, 2007; Gerasimov, & Lokazyuk, 2007; Sirodza, 2002). So, DSS may be defined as “an interactive computer system intended to support different types of activity during the decision making including poorly-structured and unstructured problems” (Tarasov, & Gerasimov, 2007; Gerasimov, & Lokazyuk, 2007).

Decision support systems contain common sets of components. The data related components are made of modules that ingest data, format data, store data, transfer data, and archive data. The algorithm related components utilize diverse methods such as atmospheric models, rule-based algorithms, fuzzy logic algorithms, statistical algorithms, the algorithm of DM under uncertainty and risk, data mining algorithms. Development of the automated systems in aviation is associated with an anthropocentric approach and on the principles of human-oriented automation, required new approaches to solving the problems of the Man-Machine interface (ICAO, 1994). One of the effective means of flight safety improvement is the inclusion of DSS to the automated systems of air traffic control (ATC) operators and pilots (Kharchenko, Shmelova, & Sikirda, 2012; 2016; Shmelova, Sikirda, & Belyaev, 2014), development DSS for flight operators (Kharchenko, Shmelova, & Sikirda, 2012; 2016). Now with the advent of UAVs, new systems are considered such as local DSSs for operators of Remotely Piloted Aircraft Systems (RPAS) (Kharchenko, Shmelova, & Bondarev, 2017).

Nowadays a new approach for flight safety provision is forming in global practice. The ICAO's term “Safety Management System” (SMS) integrates operations and technical systems with the management of financial and human resources to ensure aviation safety or the safety of the public (ICAO, 2009). Building an effective SMS of aviation activity requires integrated research of the environment in which aviation enterprises operate. Much research related to the determination of the influence of factors of internal and external management environment of aviation enterprise on the efficiency of its business. Estimating the level of influence of factors of aviation enterprise's management environment on the efficiency of SMS is a perspective task. The structural analysis of the management environment of aviation enterprise has been carried out for DSS in ANS. And, as follows from the analysis, inhomogeneous factors which influence on the aviation activity have been classified and formalized. Inhomogeneous factors of internal and external management environments of aviation enterprise have been systematically generalized using a set-theoretical approach. The influence of factors of internal and external management environment on the aviation enterprise's aviation activity has been determined (Shmelova, Sikirda, Assaul, & Stasiuk, 2015). The authors present an investigation of the management environment of aviation enterprise in ANS determined the influence of factors of internal and external management environment of aviation enterprise on aviation activity. This is a DSSs for managers in aviation enterprises. Safety management of aviation enterprise is directed on the formation and provision the achieving its goals rationally using present resources (labor, material, financial, informational, etc.). These researches are designed for aviation operators such as pilots of manned and unmanned aircraft, flight despatches, ATC operators, aviation managers, as well for scientists, engineers, air crash investigators, teachers, academicians, researchers, and students seeking current research on the application of DSS in the field of aviation.

DECISION SUPPORT SYSTEMS IN AIR NAVIGATION SYSTEM

General information About Decision Support Systems

There are many DSSs must be developed for safety assessment for each situation and stage of flight of AC. Provision of informational support of DM by aircraft operator is possible in the conditions of data collection, processing and displaying of automation within a concept of perspective global CNS / ATM system (Communication, Navigation, Surveillance / Air Traffic Management), developed by ICAO (ICAO, 2007). The “system”, as defined for the purpose of safety assessment, will always be a sub-component of some larger system, such as ANS. For example, even if the assessment encompasses all services provided within an aerodrome, this can be considered a sub-component of a larger regional system, which in turn is a subcomponent of the global aviation system. A detailed description of the system should include (ICAO, 2013):

- the purpose of the system;
- how the system will be used;
- the system’s functions;
- the system’s boundaries and the external interfaces; and
- the environment in which the system will operate.

The safety impact of a potential loss or degradation of the system will be determined, in part, by the characteristics of the operational environment in which the system will be integrated. The description of the environment should, therefore, include any factors that could have a significant effect on safety. These factors will vary from one case to another. They could include, for example, traffic characteristics, airport infrastructure, and weather-related factors. Representation of the ANS in the form of a Socio-technical System makes possible to take into account the influence of the social, cultural environment of people who DM. Taking into account the influence on DM by H-O the professional factors (knowledge, habits, skills, experience) as well as the factors of non-professional nature (individual-psychological, psycho-physiological and socio-psychological) enables to predict the H-O’s actions and flight situation development with the aid of the informational, analytic and diagnostics complex for research operator behaviour in extreme situation and to develop DSS of ANS operators for each situation and each stage of flight (Kharchenko, Shmelova, & Sikirda, 2012, 2016).

DM problems and people, who make them, recently deserved more attention. This is due to the increasing dynamism of the environment, increasing interdependence of many decisions, the high rate of scientific and technological progress. Managers who make decisions are faced with difficult choices, with the need to consider the set of alternatives. To evaluate the options the knowledge of experts, sophisticated analytical calculations, scientific research, information technologies are used. Questions of decision support at all stages of the process (goal identification, development, and DM, organization, and implementation of control) are becoming more relevant. In fact, the problem consists of the automation of the creative work of the responsible employee’s groups of organizational management – managers of all levels and those who make decisions under actual conditions of their work. Three fundamental components of DSS architecture are the database (or knowledge base) – is an information structure that reflects the status and relationship of objects analyzed; the model base – is a set of mathematical, logical, linguistic and other models used for comparative analysis of multi-alternative decision; the user interface.

Computer support of technician's activity does not require a sophisticated methodological framework and is implemented within the framework of conventional information systems. As for executives and professionals, it is important to determine the nature and characteristics of preparing and decisions making based on the level of organizational management in order to create effective software of DSS as systems which use formal rules and object management model together with the database and personal experience to develop and validate variants of management decisions. Systems of this kind do not provide the information DM but participate in it.

The main methods of analytical modelling to support DM are:

- 1) *Factor analysis* ("cause-consequence") – is changing the values of some variables (factors) and their relationships (formulae) for the investigation of values changes of other, dependent, and variable (Data Mining).
- 2) *Optimization analysis* – is getting the best value of the objective function taking into account the restrictions imposed by the selection of variables (Decision Making).
- 3) *Correlation and regression analysis* – is determining the type of relations between the dependent variables and factors (Data Science).
- 4) *Trend analysis* – is predicting the dynamics of the control object by building trends – segments of numerical series (Data Mining).

Building a DSS should ensure implementation of its important properties:

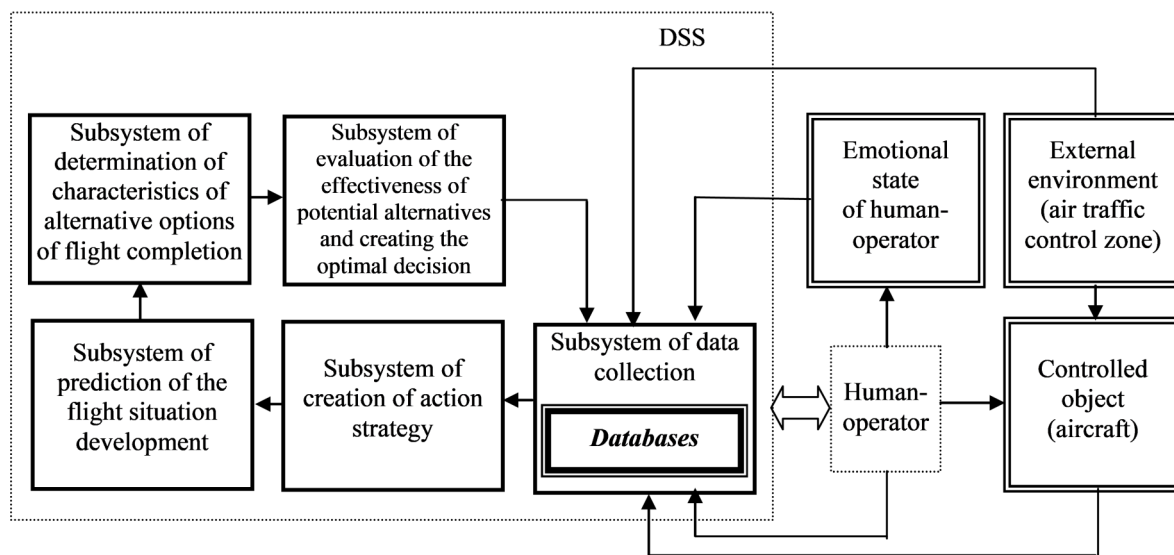
- a) *The interactivity* of DSS means that the system responds to various types of action that one intends to influence the computational process, especially in interactive mode. Human and system exchange information at a rate that is comparable with the speed of human information processing.
- b) *The integrity* of DSS – is a complex of the system component for data management and communication with users during decision support.
- c) *The power* of DSS means the ability of a system to respond to the most significant question.
- d) *Availability* of DSS – is the ability to provide answers to the user requests in the right form and at the right time.
- e) *The flexibility* of DSS describes the possibility of adapting to needs and situations changing.
- f) *The reliability* of DSS means the ability of the system to perform the required functions for a given long period.
- g) *Robustness* of DSS – is the system's ability to recover in case of error situations, both external and internal origin. For example, there are mistakes of the input data or malfunctioning hardware in a robust system.

The DSSs for H-Os in working situational have been developed (Kharchenko, Shmelova, & Sikirda, 2012, 2016; Shmelova, Sikirda, Kasatkin 2019; Shmelova, Mirskova., & Krepak, 2017; Shmelova, Konovalova, 2017; Shmelova, Sikirda, & Belyaev, 2014). Authors presented examples of DSS for aviation operators in emergency such as DSS for ATC operator in the emergency situation and DSS of UAVs operator.

Decision Support Systems of Air Traffic Controller in Emergencies

Research has shown that the choice of the optimal variant of the flight completion in emergencies, which requires AC's forced landing, requires the operator to analyze the significant amount of diverse information. For a comprehensive account of the factors that affect the DM by an air traffic controller, has been built to adaptive DSS, which allows take into account the dynamic characteristics of the state of the control object (AC) and of the external environment (ATC zone). The main tasks of ATC DSS in flight emergencies are data collection about the state of the control object (AC), external environment (ATC zone) and the individual qualities of the person, who makes the decision; creation of action strategy in flight emergency (continue flight to the destination (alternate) aerodrome or perform a forced landing); modelling of DM by H-O in flight emergency; prediction of the flight situation development during DM by H-O in flight emergency; construction of the AC's field of approachability in case of necessity of forced landing; evaluating the effectiveness of alternative decisions in flight emergency and formation of recommendations for determination of the optimal variant of flight completion. The general concept of adaptive system for informational support of the DM by air traffic controller in emergency, which requires AC's forced landing, is presented in the form of a set of subsystems (Figure 3).

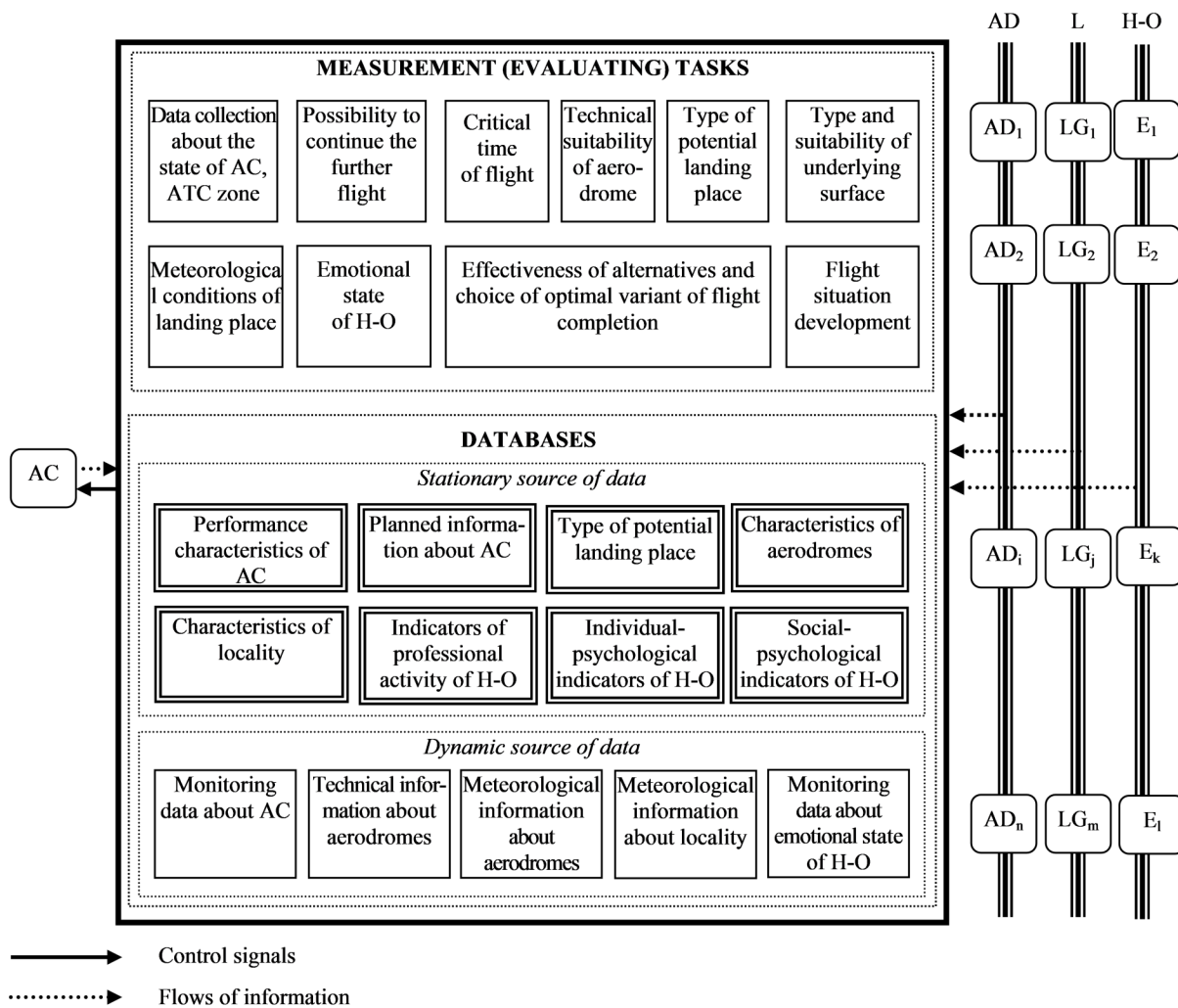
Figure 3. A conceptual model of an air traffic controller's DSS in flight emergencies, which is adaptive to changes in the state of the controlled object and of the external environment



Structure of ATC DSS in flight emergencies is shown in Figure 4. Functions of an air traffic controller's DSS in flight emergencies are expanded through information-analytical diagnostic complex for research of patterns of H-O's activity. By using complex, air traffic controller can predict the flight situation development, based on the operational monitoring of the pilot's emotional state and determining the effect of social environment on him.

The specialized program complex "Prompt" (Kharchenko, Shmelova, & Sikirda, 2012, 2016) has been created, in order to provide informational support to an air navigation system's operator. The program "Prompt" enables handling of two non-standard situations – an engine failure and fire on board of an

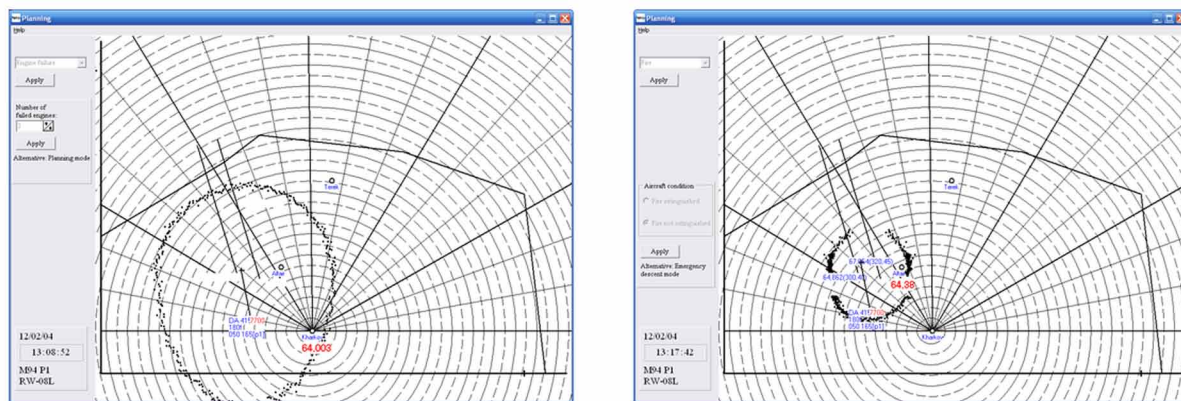
Figure 4. The structure of DSS of an air traffic controller in flight emergency, which requires AC's forced landing: AD – aerodromes, L – locality, LG – landing grounds; E – emotional state of H-O



aircraft (Shmelova, Kharchenko, Sikirda, & Gerasimenko, 2011). When the system receives a message about the unusual situation origination, the additional panel for providing the recommendations concerning the possibility of the flight continuation or necessity to execute a forced landing appears on the operator's monitor. When a type of situation – engine failure – is to be handled the number of the engines which have failed is to be input; when the type – fire – is to be handled, the condition of the aircraft is to be input and the proper advice appears on the monitor. When the recommendation concerning the necessity to accomplish a forced landing received, the field of approachability for the aircraft is built on the monitor, also values of the potential loss required for selection of a definite flight completion alternative as well as the coordinates of the potential landing sites (azimuth, range) are presented on the monitor. The program envisages the formation of the field of approachability for two extreme cases – the regime of planning and regime of an immediate descent (Figures 6). The program complex «Prompt» allows:

- to give the recommendations concerning the possibility of the further flight continuation or the necessity to accomplish a forced landing of the aircraft with the aid of the interface suitable for the user;
- to define the field of approachability of the aircraft in case when the necessity to make a forced landing arises;
- to form the evaluation of the alternative variants of the flight completion and define the optimum variant using the potential-loss minimization criterion.

Figure 5. The field of approachability and defined potential loss for a few variants of the flight completion in the event of the complete failure of the aircraft engines and in case of the fire on board the aircraft which has not been extinguished



Decision Support Systems of Unmanned Aerial Vehicle's Operator

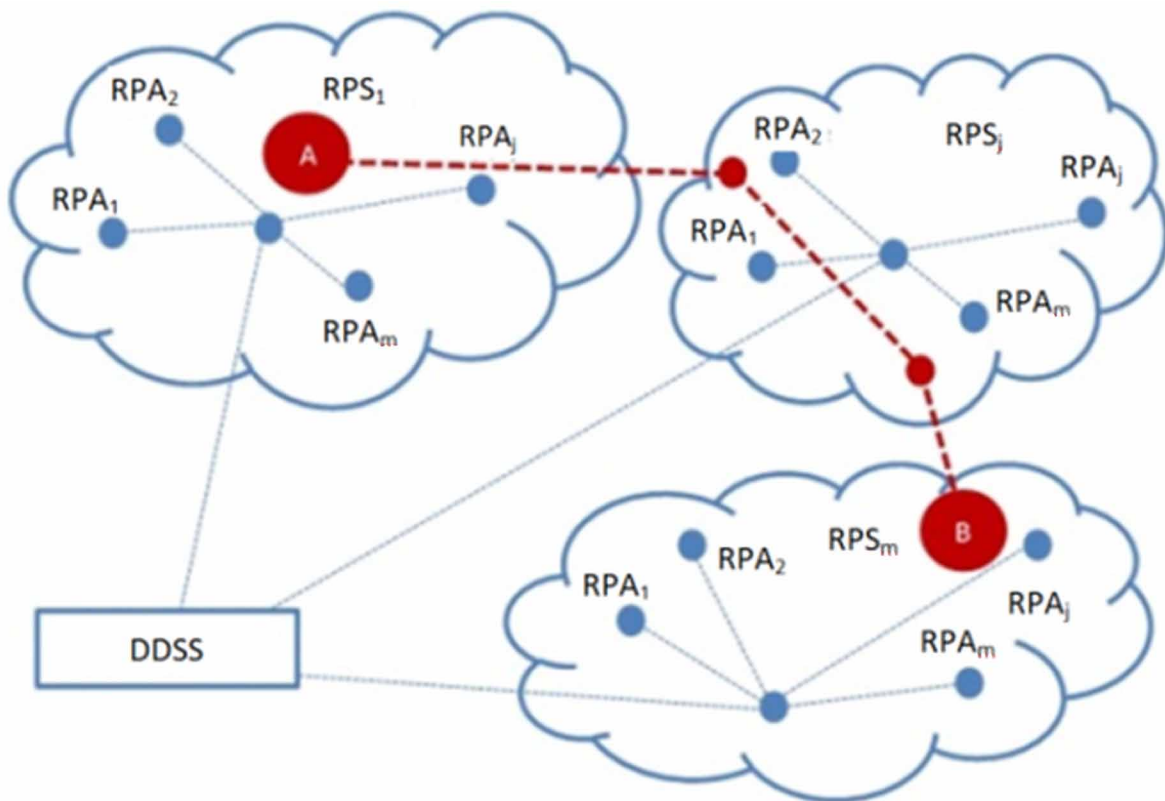
Remotely Piloted Aircraft Systems (RPAS) are a new component of the aviation system, where the ICAO, States and the industry are working to understand, define and safe integration of RPAS into non-segregated airspace. These systems are based on cutting-edge developments in aerospace technologies, offering advancements that may open new and improved civil/commercial applications as well as improvements to the safety and efficiency of all civil aviation. Unmanned aircraft has several advantages, namely low operating cost, good concealment and flexibility, simplicity and availability of technology compared to manned aircraft and Unmanned Aerial Vehicles (UAV) can be used in cases where the usage of manned aircraft is impractical, expensive or risky (Austin, 2010).

The main advantage of using UAVs is tasks that involve risk to humans and efficiency in solving economic problems. Obviously, UAVs are effective in monitoring forest fires, search and rescue operations in the processing of agricultural crops, relay communications and the movement of goods. In this sense, the usage of UAVs is more appropriate: to relay communications at places where the antenna coverage cannot be set because of difficult terrain, in agriculture, aerial photography, moving cargo, for military purposes. Emergency situations may occur when flying both in manual and in the autonomous management. For operations carried out “manually” plays an important role in the human factor and a significant part of emergency arises due to the wrong actions of the operator. Today implementation of RPAS is necessary because they have transformed the idea of airpower in modern war times; they are smaller and less expensive than jet aircraft; don't put the pilots at risk when crash. Civil aviation has, to

this point, been based on the notion of a pilot operating the aircraft from within the aircraft itself and more often than without passengers on board. Removing the pilot from the aircraft raises important technical and operational issues, the extent of which is being actively studied by the aviation community. Most importantly, the introduction of remotely piloted aircraft into non-segregated airspace and at aerodromes should in no way increase safety risks to manned aircraft (ICAO, 2015).

Global operational air traffic management concept that reflects the ICAO vision of a single coherent and based on global modern mutually ANS systems should ensure collective consolidation to traffic management. Consolidation of information flows plays an important role in air traffic management. To make the right decision information that comes for air traffic controllers should be structured and consolidated information flows. In the implementation of this task the choice of database plays an important role. For rational aggregating requests and increase the processing speed NoSQL database has been proposed (Shmelova, 2016; Shmelova T., Sterenharz A., Burlaka O., 2019). Distributed DSS (DDSS) represents a complex system with complex interactions geographically distributed local DDSS operators of Remotely Piloted Aircraft (RPA). RPA is controlled by Remote Pilot Station (RPS) with the Command-and-Control (C2) line (Figure 6). Together with other components such as the starter equipment and equipment for the return, if it is used, RPA, RPS, and the C2 line constitute RPAS (ICAO, 2015).

Figure 6. The structure of distributed RPS Mission Control UAVs



The operational procedures of RPAS are determined by the aim of flight, rules of flights, areas of flights and functional data link from RPA. As the technologies develop RPA make their transition from strictly military usage to civil aviation and currently serve to fulfil many civil purposes. Actually it's very interesting to use DSSs for decision problems of pre-flight preparation, flight planning, DM in emergency situations. The main rule of effective realization of flight for piloted and unmanned aviation too is mandatory performance of pre-flight preparation including review and maintenance of the RPAS and RPS, obligatory checking of communication equipment and handling functions, configuration of RPS, downloading and checking of information that relates to flight planning procedure. Additionally, if it is necessary the clearance from ATC organizations should be received. One of the tasks that solve RPA pilot in the process of pre-flight preparation is a selection of alternate aerodromes/places for return procedure in the case of emergency situation that is caused by inappropriate meteorological conditions. Following documents regulate the selection of alternate aerodrome: flight manual, "Operation of decision making during IFR take-off and landing of aircraft of civil aviation of Ukraine" and other regulatory documents. The usage of RPA should be regulated by the same regulatory documents as for ordinary aircraft. RPA pilot should take into account the following factors:

- adequacy of fuel/energy reserve;
- distance to the alternate aerodrome/place;
- reliability of C2 lines for connection with RPA;
- possibility of communication with ATC units;
- meteorological conditions at the alternate aerodrome/place.

The multifunctional model of selection of alternate aerodrome/place of RPA has been proposed for economical effectiveness of flight realization of RPA which is used in remote system of decision making support of UAV's operator.

Illustrative Examples of Building Models for DSSs of UAV's Operator

To show the DSSs we use individual works of aviation students in education: Shulimov Oleh (*Example 1*), Fomin Olexandr (*Example 2*), (course "Basic of DM in ANS" in National Aviation University, Kyiv). In these examples the decision of selection tasks of alternate aerodrome/place in the case of emergency landing (difficult meteorological conditions, etc.) by method of DM in uncertainty is means of the criteria of DM in uncertainty: Wald, Laplace, Savage, Hurwicz. During approach of RPA of type N that flies along the route from the point A to point B , deterioration of meteorological conditions on the destination aerodrome appears. The task is to find optimal aerodrome/place of landing in the case of regular and irregular flight (first time realization).

Example 1: In a Case of Big UAV

Let's calculate the optimal landing aerodrome/place for RPA flying from Kharkiv to Kiev (Borispol) in a case of bad weather conditions: flight regular; for the first time; after two weeks. The matrix of possible results of DM that relates to choosing of optimal landing aerodrome/place in Table 1 has been obtained using MS EXCEL.

Table 1. The matrix of possible results of decision making that relates to choosing of landing aerodrome/ place for example 1

Alternative decisions		Factors that influence on decision making						Solutions			
A_i	A_{AAs}	Availability of fuel/energy onboard of RPA	Distance from RPA to $A_{Dest}, A_{Dep}, A_{AP}$	TTC of $A_{Dest}, A_{Dep}, A_{AP}$	Meteorological conditions at $A_{Dest}, A_{Dep}, A_{AP}$	Reliability of C2 lines for connection with RPA	Possibility of communication with ATC units	W	L	H	S
A_1	A_{ADest}	1	1	9	6	5	7	1	4,83	5	8
A_2	A_{ADep}	8	9	9	2	8	8	2	7,33	5,5	7
A_3	A_{AAP1}	5	6	3	3	6	3	3	4,33	4,5	3
A_4	A_{AAP2}	6	5	2	5	5	2	2	4,17	4	6
A_5	A_{AAP3}	5	7	7	4	4	1	1	4,7	4	6

Thus, according to the Wald and Savage (after calculation using the regret matrix) criteria the optimal solution is the alternative A3; according to the Laplace and Hurwitz criteria the optimal solution is the alternative A2.

Example 2: In a Case of Small UAV

As an example for our task we are going to consider RPA flight from Bila Tserkva aerodrome to Konotop with possible alternate destinations at Vasylkiv, Berezan' Nizhyn and Pryluky. For each aerodrome we have introduced risk parameters and corresponding coefficient which are going to serve as a basis for our investigation. The following parameters have been introduced: remoteness of the aerodromes, availability of fuel on-board, TTC of runways, reliability of C2 links, lightning systems and approach systems at aerodromes, meteorological conditions and navigational aids (Tables 2).

Table 2. The matrix of possible results of decision making that relates to choosing of landing aerodrome/ place for example 2

Alternative decisions		Factors that influence on decision making							Solutions			
A_i	A_{AAs}	Remoteness of the aerodromes	Availability of fuel onboard	TTC of runways	Reliability of C2 links	Lighting systems	Meteorological conditions	Navigational aids	W	L	H	S
A_1	Bila Tserkva	9	2	5	8	0	3	9	0	5,14	4,5	9
A_2	Konotop	3	5	7	9	2	4	9	2	5,57	4,5	7
A_3	Vasylkiv	2	8	8	9	2	4	10	2	6,14	6	8
A_4	Berezan'	7	1	8	7	1	7	7	1	5,43	5	7
A_5	Nizhyn	6	4	8	6	6	5	8	4	6,14	4,5	4
A_6	Pryluky	4	8	9	8	4	6	6	4	6,43	6,5	5

Thus, according to the Wald criterion, the optimal solutions are the alternatives A5 and A6; according to the Laplace and Hurwitz criteria the optimal solution is the alternative A6; according to the Savage (after calculation using the regret matrix) criterion the optimal solution is the alternative A5. For decision making using four criterion of estimation, such as Wald, Laplace, Hurwicz and Savage, the computer program “Classic Decision Criteria: Wald, Laplace, Hurwitz, Savage” (Shmelova, Shulimov, Chorna, & Kovtunets, 2015) has been designed.

FUTURE RESEARCH DIRECTIONS

Further research should be directed to the solution of the problem in prerequisites of emergency situations and preventing catastrophic situations too. Models of flight emergency development and of DM by ANS's H-O in-flight emergency will allow predicting the operator's actions with the aid of the informational-analytic and diagnostics complex for research of H-O behavior in extreme situations. It is necessary to develop collaborative DSSs of H-O of ANS (pilots, ATCs, flight dispatchers, UAV and SMS operators) for consolidation DM in an emergency. In future develop DSSs for operators in other areas such as chemical production, energy, military industry, medicine, etc.

On the next steps we are planning consider other collaboration DSSs, such as for hydraulic engineering, chemical and military industries, gas and oil pipelines, nuclear power plants and transport, etc. Developing of AI systems, Intelligent DSSs considering new concepts in aviation (PBA, CDM, SWIM, SMART, AI, etc.) for different operators and each stage, process, which are problems, with using modern information technologies Data Science, Big Data, Data Mining, DM, etc. It is necessary to analyze the all factors influencing on the DM of operators in these systems in order to predict the development of the technogenic catastrophe and prevent it with using Intelligent DSSs.

CONCLUSION

Conceptual models of air traffic controllers DSS in-flight emergencies and of flight dispatcher's DSS have been presented, program realization of systems has been shown. DSS of air traffic controller in-flight emergencies gives to the aviation operator possibility to quantify evaluation of alternatives of the flight situation development and to operatively choose an action strategy with minimal potential damage in the conditions of incomplete and uncertainty information. DSS of flight dispatcher provides the fast search of the information and support of the aviation operator during DM regarding the possibility of departure or regarding the optimal aerodrome in case of the forced landing of AC. It is reasonable to use an air traffic controllers DSS during the real Air Traffic Control in-flight emergencies, a flight dispatcher's DSS – in the automated system of pre-flight information preparation. The structure of the distributed DSS for RPA has been developed. For rational aggregating requests and increase the processing speed NoSQL database has been proposed. UAV (RPA) during the flight can be controlled by RPAS. RPA is controlled by RPS with C2 line. The computer program for the finding optimal landing aerodrome/place in an emergency situation on board of RPA that is caused by meteorological conditions has been designed.

REFERENCES

- Austin, R. (2010). *Unmanned Aircraft Systems: UAVs Design, Development and Deployment*. John Wiley & Sons Ltd. doi:10.1002/9780470664797
- Federal Aviation Administration. (2008). Aeronautical Decision Making. In Pilot's Handbook of Aeronautical Knowledge. FAA-H-8083-25. 17.1–17.32.
- Gerasimov, B., & Lokazyuk, V. (2007). *Intelligent Decision Support Systems: monograph*. European University Press.
- IATA. (n.d.). <https://www.iata.org/publications/Pages/AI-white-paper.aspx>International
- International Civil Aviation Organization. (1994). Human Factors in CNS/ATM Systems. Introduces the Concept of Human-Centred Automation: Human Factors Digest N° 11. Circ. ICAO 249-AN/149. Canada, Montreal: Author.
- International Civil Aviation Organization. (2004). Cross-Cultural Factors in Aviation Safety: Human Factors Digest N° 16. Circ. ICAO 302-AN/175. Canada, Montreal: Author.
- International Civil Aviation Organization. (2005a). Global Air Navigation Plan. Doc. ICAO 9750-AN/963. Canada, Montreal: Author.
- International Civil Aviation Organization. (2005b). Global Air Traffic Management Operational Concept. Doc. ICAO 9854-AN/458. Canada, Montreal: Author.
- International Civil Aviation Organization. (2008). Normal operations safety survey (NOSS) (1st ed.). Doc. ICAO 9910-AN/473. Canada, Montreal: Author.
- International Civil Aviation Organization. (2009). Manual on Global Performance of the Air Navigation System. Doc. ICAO 9883. Canada, Montreal: Author.
- International Civil Aviation Organization. (2012). *Report of the Twelfth Air Navigation Conference*. AN-Conf/12-WP/162. Canada, Montreal: Author.
- International Civil Aviation Organization. (2013). Safety Management Manual (SMM) (3rd ed.). Doc. ICAO 9859-AN/474. Canada, Montreal: Author.
- International Civil Aviation Organization. (2013). *State of Global Aviation Safety*. Author.
- International Civil Aviation Organization. (2015). Manual of Remoted Piloted Aircraft Systems (RPAS). Doc. ICAO 10019-AN/507. Canada, Montreal: Author.
- International Civil Aviation Organization. (2017). *Global Aviation Security Plan (GASP)*. International Civil Aviation Organization.
- International Civil Aviation Organization. (2017). *Global Air Navigation Plan (GANP, Doc 9750)*. Author.
- International Civil Aviation Organization. (2018). *Potential of Artificial Intelligence (AI) in Air Traffic Management (ATM)*. In Thirteenth Air Navigation Conference ICAO, Montréal, Canada.
- Kharchenko, V., Shmelova, T., Bondarev, D., & Stratiy, A. (2017). Modelling of Decision Making of Unmanned Aerial Vehicle's Operator in Emergency Situations. *Proceedings of the National Aviation University, 1*(1), 20–28. doi:10.18372/2306-1472.70.11419

Kharchenko, V., Shmelova, T., & Sikirda, Y. (2012). *Decision-Making of Operator in Air Navigation System*: monograph. Kirovograd: KFA of NAU.

Kharchenko, V., Shmelova, T., & Sikirda, Y. (2016). *Decision-Making in Socio-Technical Systems*. Kyiv: NAU.

Shmelova, Scarponi, & Chialastri. (2018). Deterministic and Stochastic Models of Decision Making in Air Navigation Socio-Technical System. *Proceedings of the International Conference on ICT in Education, Research and Industrial Applications*.

Shmelova, T., & Bondarev, D. (2016). Graph Theory Applying for Quantitative Estimation of UAV's Group Flight into Aerial Photography. *Electronics and Control Systems*, 4(46).

Shmelova, T., Bondarev, D., & Znakovska, Y. (2016). Modeling of the Decision Making by UAV's Operator in Emergency Situations. In *4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC-2016)*. Kyiv: NAU. 10.1109/MSNMC.2016.7783099

Shmelova, T., Kharchenko, V., Artemenko, O., & Otryazhiy, V. (2011). *Computer Program "Choosing the Pre-Flight Information and Decision Making about Departure for Automated System of Pre-Flight Information Preparation (AS PIP)": certificate of registration of copyright on the product N° 40062*. Academic Press.

Shmelova, T., Kharchenko, V., Sikirda, Y., & Gerasimenko, O. (2011). *Computer Program "Optimizing the Choice of Alternative Variant of Aircraft Flight Completion in Flight Emergencies "Prompt": certificate of registration of copyright on the product N° 3787*. Academic Press.

Shmelova, T., & Konovalova, A. (2017) Decision making of ATCO in emergency situation "aircraft decompression. *Proceedings of the International conference "AVIA -2017"*.

Shmelova, T., Mirskova, M., & Krepak, D. (2017) Decision making in emergency situation of air navigation system's operator. *Proceedings of the International conference "AVIA -2017"*.

Shmelova, T., Shulimov O., Chorna M., & Kovtunets, O. (2015). *Computer Program "Classic Decision Criteria: Wald, Laplace, Hurwitz, Savage": certificate of registration of copyright on the product N° 60624*. Academic Press.

Shmelova, T., Sikirda, Y., Assaul, A., & Stasiuk, O. (2015). Structural Analysis of Management Environment of Aviation Enterprise from the Point of Systematic Approach. *Proceedings of the National Aviation University*, 2(2), 3–12. doi:10.18372/2306-1472.63.8828

Shmelova, T., Sikirda, Y., & Belyaev, Y. (2014). Informational Support of Air Navigation System's Human-Operator. *Scientific Works of the National University of Food Technologies*, 20(4), 7–18.

Shmelova, T., Sikirda, Y., & Kasatkin, M. (2019). Applied Artificial Intelligence for Air Navigation Sociotechnical System Development. *Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer*.

Shmelova, T., Sterenharz, A., & Burlaka, O. (2019) Optimization of flows and flexible redistribution of autonomous UAV routes in multilevel airspace. *5th International Workshop on Theory of Reliability and Markov Modelling for Information Technologies (TheRMIT)*, 4.

Shmelova, T., & Stratij, A. (2015). *Computer Program “Aggregation of Heterogeneous Information Flows”*: certificate of registration of copyright on the product N° 59466. Academic Press.

Shmelova, T., Yakunina, I., & Moiseenko, V. (2014). *Computer Program “Network Analysis of a Special Case in Flight”*: certificate of registration of copyright on the product N° 55587. Academic Press.

Shmelova, T., Yakunina, I., & Yakunin, R. (2013). *Computer Program “TestExtraordinaryIncident”*: certificate of registration of copyright on the product N° 50826. Academic Press.

Shmelova, T., Yakunina, I., Znakovskaya, E., & Ivleva, G. (2014). *Computer Program “Decision Making under Uncertainty”*: certificate of registration of copyright on the product N° 56819. Academic Press.

Shmelova, T., & Znakovskaya, E. (2015). *Computer Program “Decision Tree”*: certificate of registration of copyright on the product N° 62163. Academic Press.

Sikirda, Yu., Shmelova, T., & Tkachenko, D. Automated System for Evaluation of the Organizational Risk Factors Influence on Flight Safety in Air Traffic Control *5th International Conference on Methods and Systems of Navigation and Motion Control*. 10.1109/MSNMC.2018.8576317

Sirodzha, I. (2002). *Quantum Models and Methods of Artificial Intelligence for Decision Making and Management*. Naukova Dumka.

Tarasov, V., & Gerasimov, B. (2007). *Intelligent Decision Support Systems: Theory, Synthesis, Efficiency*. Kyiv: International Academy of Computer Science and Systems.

KEY TERMS AND DEFINITIONS

AC: Aircraft.

ADS-B: Automatic dependent surveillance-broadcast.

Air Traffic Management (ATM): Is an aviation term encompassing all systems that assist aircraft to depart from an aerodrome, transit airspace, and land at a destination aerodrome, including Air Traffic Services (ATS), Airspace Management (ASM), and Air Traffic Flow and Capacity Management (ATFCM).

ANS: Air navigation system.

AS PIP: Automated system of pre-flight information preparation.

ASSIST: Acknowledge, separate, silence, inform, support, time.

ASSSIST: Acknowledge, separate, synergetic (coordinated, cooperation, consolidation), silence, inform, support, time.

ATC: Air traffic controller.

BVLOS: Beyond visual line of sight.

CNS/ATM: Communication, navigation, surveillance/air traffic management.

Collaborative Decision Making (CDM): Is a joint government/industry initiative aimed at improving air traffic flow management through increased information exchange among aviation community stakeholders.

Command and Control (C2): Denotes the set of organizational and technical attributes and processes by which an enterprise marshals and employs human, physical, and information resources to solve problems and accomplish missions.

Communication, Navigation, and Surveillance (CNS): Are the main functions that form the infrastructure for Air Traffic Management, and ensure that air traffic is safe and efficient.

Decision Support System (DSS): Is the interactive computer system intended to support different types of activity during the decision making including poorly-structured and unstructured problems.

Distributed Decision Support System (DDSS): Is a decision support system which supports distributed organizational decision making.

DM: Decision making.

Flight and Flow Information for a Collaborative Environment (FF-ICE): A product of the ICAO Global ATM Concept, that defines information requirements for flight planning, flow management and trajectory management and aims to be a cornerstone of the performance-based air navigation system.

Global Navigation Satellite Systems (GNSS): Is the common name of global satellite-based systems for navigation, positioning and time transfer. The most well-known system is the American system GPS (Global Positioning System) which got its operational status in 1994. Russia also has a similar system since the beginning of the 1990's called GLONASS (GLObal NAVigation Satellite System).

H-O: Human-operator.

ICAO: International Civil Aviation Organization.

IFR: Instrument flight rules.

NoSQL (Originally Referring to “Non-SQL” or “Non-Relational”) Database: Provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

PANS: Procedures for air navigation services.

Performance-Based Approach (PBA): Is a decision-making method based on three principles: strong focus on desired/required results; informed decision making driven by those desired/required results and reliance on facts and data for decision making. The PBA is a way of organizing the performance management process.

Remotely Piloted Aircraft (RPA)/Remotely Piloted Aircraft System (RPAS)/Unmanned Aerial Vehicle (UAV)/Drone: Is a system that based on cutting-edge developments in aerospace technologies, offering advancements which are opening new and enhanced civil-commercial applications as well as improvements to the safety and efficiency of the entire civil aviation. The terms RPA or UAV are used to describe the aircraft itself, whereas the term RPAS is generally used to describe the entire operating equipment including the aircraft, the control station from where the aircraft is operated and the wireless data link.

RPS: Remote pilot station.

Safety Management System (SMS): Is the formal, top-down, organization-wide approach to managing safety risk and assuring the effectiveness of safety risk controls. It includes systematic procedures, practices, and policies for the management of safety risk.

SARPs: Standards and recommended practices.

SMART: Is a best practice framework for setting goals. A SMART goal should be specific, measurable, achievable, relevant and time-bound.

System-Wide Information Sharing and Management (SWIM): Is a Federal Aviation Administration advanced technology program designed to facilitate greater sharing of Air Traffic Management system information, such as airport operational status, weather information, flight data, status of special use airspace, and National Airspace System restrictions.

TTC: Tactical and technical characteristics.

VFR: Visual flight rules.

VLOS: Visual line of sight.

Classification and Recommendation With Data Streams

5

Bruno Veloso <https://orcid.org/0000-0001-7980-0972>*INESC TEC, Portugal & University Portucalense, Portugal***João Gama** <https://orcid.org/0000-0003-3357-1195>*INESC TEC, Portugal & FEP, University of Porto, Portugal***Benedita Malheiro** <https://orcid.org/0000-0001-9083-4292>*Polytechnic Institute of Porto, Portugal & INESC TEC, Portugal*

INTRODUCTION

Internet of Things (IoT), wireless sensor networks, social networks, crowdsourcing platforms and personal mobile devices are some examples of data stream sources. They continuously generate large volumes of real time data at variable rates, making traditional off-line processing obsolete. Data streams are sequences of timely ordered interactions between members of dynamic collections (Hopfgartner, Kille, Heintz, & Turrin, 2015). While continuous flows of information, they need real time incremental processing techniques to extract meaningful and timely information. With the exponential growth of data stream sources, data stream processing has become indispensable. Data stream algorithms have three main challenges: (i) present lower complexity to keep up with the data rate arrival and good scalability; (ii) display a manageable memory footprint; and (iii) adopt incremental techniques to build and maintain the model(s) updated. Specifically, the model updating process needs to prioritise information, *e.g.*, using novelty and relevance criteria, as well as to identify relevant data changes. Such challenges and problems require the design and development of new techniques to balance the computational complexity and the prediction performance. Data stream processing techniques have been extensively covered in several data mining books (Muthukrishnan, 2005), (Aggarwal, 2007), (Gama & Gaber, 2007), (Bifet, 2010), (Galić, 2016) (Garofalakis, Gehrke, & Rastogi, 2016). This article addresses data stream knowledge processing, ranging from classification to recommendation and evaluation, describing existing techniques, typical fields of application and identifying open challenges.

BACKGROUND

Data stream processing typically includes classification, recommendation and evaluation. Data stream classifiers work incrementally and attempt to classify events as they occur, *i.e.*, in near real time and not *a posteriori*. Micro-clusters (Aggarwal, Han, Wang, & Yu, 2004), decision trees (Rutkowski, Pietruczuk, Duda, & Jaworski, 2013), ensemble classifiers (Osojnik, Panov, & Dzeroski, 2017) and adaptive model

DOI: 10.4018/978-1-7998-3479-3.ch047

rules (Duarte, Gama, & Bifet, 2016) are well-known classification techniques which have been adapted to work with data streams. However, they suffer from concept drift, data outliers and missing data. Concept drifts can occur with time, *e.g.*, when interests change, and may arise from different situations, *e.g.*, changes in the properties of the data or inappropriate hyper-parameter tuning, are identifiable by alterations in the statistical properties of the data. In such cases, past observations become irrelevant to the current state and the algorithm needs to forget to improve its accuracy. Outliers correspond to data spatially distant from the trend of most of the observations. Outlier observations can be considered as hidden variables, inducing large statistical errors, and may indicate experimental error or unexpected variability in the measurement or noise. Missing values result from sensor malfunctioning or communication problems.

The data stream recommendation engines found in the literature rely on content-based or collaborative filters to generate personalised recommendations and popularity, recency, randomness or trending filters to make general recommendations. Popularity filters order items based on the item selection frequency / average rating; recency filters order the items based on the arrival date; the trend-based filters compute the frequency or rating trend of the items within a temporal sliding window; and random filters select randomly items to create a serendipity effect. To make personalised recommendations, content-based filters compute the similarity between the items previously selected and those not yet chosen by the active user, whereas collaborative filters learn user preferences and chose items appreciated by users with interests identical to the active user. With exception of collaborative filtering, the remaining approaches do not incorporate learning, which is essential to be efficient and reactive to the different user behaviours. Consequently, many of the simpler general filters tend to be used in combination with collaborative filtering. For this reason, we will focus our analysis on real time collaborative filters, which are extended versions of standard off-line model-based collaborative filters. They adopt incremental matrix factorisation (Peterek, 2007), (Takács, Pilászy, Németh, & Tikk, 2009), (Vinagre, Jorge, & Gama, 2014) to decompose the large user-item matrix into a product of smaller matrices and, then, apply other mathematical operations such as gradient based optimisation. By default, this technique has multiple challenges such as scalability, cold start, hyper-parameter optimisation and data relevance.

Finally, the evaluation of data stream algorithms, *i.e.*, the assessment of the quality of the predictions generated by a data stream classifier or recommendation engine, requires dedicated techniques such as sliding windows (Li, Maier, Tufte, Papadimos, & Tucker, 2005) (Ghanem, Hammad, Mokbel, Aref, & Elmagarmid, 2007), the prequential approach presented by (Gama, Sebastião, & Rodrigues, 2009) and the prequential Area Under the Curve (AUC) (Brzezinski, & Stefanowski, 2017).

LIMITATIONS OF EXISTING CLASSIFICATION AND RECOMMENDATION ALGORITHMS

The stream-based recommendation and classification algorithms present limitations in terms of data representation, model maintenance, performance and evaluation. Typically, researchers develop and tune their algorithms for off-line processing, using stored data sets and representing all entities. However, this approach is incompatible with on-line real-world data stream processing applications where events may arrive at various rates and tend to accumulate rapidly, generating large volumes of data. Furthermore, data streams, due to their inherent properties, are highly prone to data variability or concept drift, outliers and missing data. To timely process all events received, not only the incremental data processing has to remain simple, *i.e.*, have lower complexity, but parsimonious data representation techniques must be

adopted. In the latter case, it is necessary to prioritise data in terms of relevance. The goal is to minimise the memory footprint by discarding less relevant, outdated or noisy data.

The proper assessment of the stream-based algorithms raises other problems. With these algorithms, it is impossible to adopt standard off-line evaluation protocols, such as holdout or cross validation, since they are applied to the complete data set, which, in the case of a data stream, is by definition unknown. To overcome this problem, it's necessary to adopt incremental evaluation metrics or sliding windows to measure the performance of the algorithm considering the available data.

STREAM-BASED CLASSIFICATION, RECOMMENDATION AND EVALUATION

The focus of the article is to provide an overview of classification and recommendation with data streams. Specifically, it will cover existing techniques and report on problems, successes and open challenges.

In terms of online classifiers, the most used techniques are the micro clusters, decision trees, ensemble classifiers and adaptive model rules. However, the classification algorithms suffer from three types of problems: concept drift, which is a significant change on the statistical properties of the data, the occurrence of outliers, which introduces noise in the models and reduces the accuracy of the predictions, and the appearance of unlabelled data, which can be minimised by using imputation data.

The online recommendation algorithms typically use matrix factorisation techniques to build a preference model. These models can be extended using additional feature information to build more accurate models, *e.g.*, Anyosa, Vinagre & Jorge (2018) adopted a co-factorisation technique to add additional features to the model. However, these algorithms store all user preferences, including old user preferences which may no longer represent the current user tendencies. To minimise this drawback, relevance-based techniques, *e.g.*, forgetting, can be applied.

In terms of evaluation for online prediction algorithms, it was necessary to define new incremental performance measures to assess the performance of the prediction algorithms, *e.g.*, the predictive sequential evaluation protocol (Gama, Sebastião, & Rodrigues, 2009) and the usage of sliding windows to measure with more accuracy the learning behaviour of the data stream algorithms (Datar, Gionis, Indyk, & Motwani, 2002).

CLASSIFICATION

Data stream incremental classification techniques include: (i) micro-clusters, representing classes, which are built from clusters of data stream instances together with the corresponding summary of the statistics (Aggarwal, Han, Wang, & Yu, 2004); (ii) decision tree graph structures, holding in each node the available decisions and possible consequences (Rutkowski, Pietruczuk, Duda, & Jaworski, 2013); (iii) ensemble learning methods, which divide the problem into sub-problems, by applying individual classifiers specialized on each sub-problem, and combine the predictions of the individual classifiers to classify the incoming instances (Krawczyk, Minku, Gama, Stefanowski & Woźniak, 2017); and (iv) adaptive model rules, where data is split into several models or classes to build more accurate local models by creating new rules and discarding old rules (Duarte, Gama, & Bifet, 2016). On-line classifiers perform incremental adjustments to update the set of classes using different techniques: (i) gradient descent in the case of adaptive model rules (Duarte, Gama, & Bifet, 2016); (ii) recursive replacement of leaves by nodes, which verify the test conditions, in the case of decision trees (Hulten, Spencer,

& Domingos, 2001); (iii) random forest walks, an ensemble learning method which creates multiple alternative decision tree models to produce the best predictions together with resampling and adaptive operators to overcome concept drifts (Gomes et al., 2018); and (iv) ensemble learning with multi-target regression trees to classify multi-labels (Osojnik, Panov, & Dzeroski, 2017).

Several research efforts have been done to overcome concept drift, outliers and missing data affecting classification. Klinkenberg (2001) exploit the use of unlabelled data and sliding windows to discard irrelevant data. Hulten, Spencer, & Domingos (2001) adopt decision trees with a sampling window to improve the accuracy of the model without sacrificing the processing speed. Wang, Fan, Yu, & Han (2003) propose the use of ensemble classifiers to minimise the problem of the concept drift and outliers. Angiulli & Fassetti (2007) detected outliers based on the distance between an object and its neighbours, using the Euclidian distance. Pokrajac, Lazarevic & Latecki (2007) describe an outlier detection scheme based on the distance between observations and their density. Jaworski, Duda, Rutkowski, Najgebauer, & Pawlak (2018) suggest the adoption of Parzen kernels – a regression technique – extended with sliding windows and forgetting factors to detect concept drifting. To surpass the missing data problem, Saar-Tsechansky, & Provost (2007) suggest the use of imputation, omission or analysis. Specifically, Montiel, Read, Bifet, & Abdesslem (2018) automatically impute missing data, regardless of the characteristics of the data set, by means of a cascade imputation technique.

Examples of applications of this type of algorithms are sensor monitoring, financial and marketing predictive models.

RECOMMENDATION

Recommendation algorithms for data streams are incremental. There are several approaches in the literature, ranging from matrix factorization (Koren, Bell, & Volinsky, 2009), matrix factorization with gradient descent (Burgess et al., 2005) to ensemble methods (Vinagre, Jorge & Gama, 2018) and co-factorisation (Anyosa, Vinagre, & Jorge, 2018). Every time a new user event arrives, the model is updated, and new personalised recommendations are provided. Matrix factorisation scalability is an issue due to the continuous arrival of new users and items, producing huge sparse matrices, which require large volumes of memory. To overcome this scalability problem of collaborative filters, it is possible to partition the data or migrate to the cloud, *e.g.*, using a distributed algorithm like Apache Spark together with Hadoop (Meng et al., 2016).

The representation of new users and items – cold-start – is particularly hard to overcome in the case of data streams. A typical approach is to add new lines or columns to the matrix model and initialise the new latent vectors, *e.g.*, with random values centred around a predefined value (Matuszyk, Vinagre, Spiliopoulou, Jorge, & Gama, 2017) or the average value of similar items.

Standard hyper-parameter optimisation is normally performed off-line, involving multiple training and validation stages, and include grid search, random search, stochastic gradient or Bayesian optimisers. The only exception are the works of Pfaffe, Tillmann, Walter, & Tichy (2017) and Veloso, Gama, & Malheiro (2018), where the Nelder-Mead optimisation is adopted to tune model hyper-parameters on-line.

Another important aspect when processing data streams is the ability to recognise the relevance of the data, *e.g.*, to distinguish between old and new information. However, since matrix factorisation algorithms build models based on the all observations received so far (Vinagre, Jorge & Gama, 2014), they are unable to discriminate between old and recent data. Matuszyk et al. (2017) proposed the adoption of factor fading together with first in, first out (FIFO) queues and (Koren, 2010) introduced the usage

of time-dependent bias. Anyosa, Vinagre & Jorge (2018) explore co-factorisation algorithms to build a more accurate models, taking into consideration additional features. Wang et al. (2018) propose a probability-based approach by combining SGD with pairwise probabilistic matrix factorisation models.

More recently, there is a trend to adopt ensemble methods for data stream recommendation data to outperform single models. Vinagre, Jorge, & Gama (2018) apply ensemble methods to reduce the variance of data and Matuszyk, & Spiliopoulou (2018) rely on specialised ensemble semi-supervised co-trainers to alleviate the problem of data sparsity.

These algorithms are extensively used within e-commerce, video and music streaming platforms.

EVALUATION

The typical data stream evaluation metrics determine the error, *e.g.*, the incremental Root Mean Square Error (RMSE) (Takács, Pilászy, Németh, & Tikk, 2009), and the accuracy, *e.g.*, the incremental Recall@ N (Cremonesi, Koren, & Turrin, 2010). The incremental RMSE calculates the RMSE after each new viewer rating event (Takács, Pilászy, Németh, & Tikk, 2009). The incremental Recall@ N , for each new event, selects 1000 items not yet rated by the active user, adds the newly rated item and, then, makes predictions for this subset of 1001 items. Finally, it sorts the 1001 predictions by descending order and, if the newly rated item belongs to the list of the top N viewer predictions, counts a hit (Cremonesi, Koren, & Turrin, 2010).

Regarding the evaluation of data streams, there are three main approaches: (i) sliding windows to measure the performance metrics in given timespan (Datar, Gionis, Indyk, & Motwani, 2002); (ii) predictive sequential evaluation to calculate the performance metrics incrementally (Gama, Sebastião, & Rodrigues, 2009), *i.e.*, instance by instance; and (iii) the prequential Area Under the Curve (AUC) (Brzezinski, & Stefanowski, 2017). Typically, sliding windows are data windows with predefined size and slide rate where the evaluation metrics are calculated. The predictive sequential evaluation – prequential evaluation – assesses the predictions generated for each new event before updating the model, using the incremental RMSE and Recall@ N . The prequential AUC, suggested by Brzezinski, & Stefanowski (2017), uses a sorted tree and a forgetting sliding window to compute the AUC as originally proposed by (Hanley & McNeil, 1982). Depending on the data rate, the evaluation process may still take too long, leading to the adoption sample-based evaluation methods.

According to (Krawczyk, Minku, Gama, Stefanowski, & Woźniak, 2017), apart from the predictive accuracy or error, additional performance metrics should be monitored and taken into account during properly executed evaluation of streaming algorithms: memory consumption, update time and decision time. The memory footprint of stream algorithms needs to be monitored to know beforehand the type and quantity of resources required. The model update time provides an indication of the complexity and efficiency of the adopted algorithm (the update time should be lower than the arrival interval between instances). The decision time corresponds to the time interval required by the model to make a decision regarding new events. This is especially crucial for algorithms that cannot update and make predictions regarding new instances at the same time.

Last, but not least, the evaluation of data stream algorithms tends to be non-reproducible due to the endless nature and data volume. To avoid this problem, researchers create partial data sets from real data streams or reuse existing data sets with timestamped data for benchmarking data stream algorithms.

FUTURE RESEARCH DIRECTIONS

Although data stream processing is a highly active research domain, there are several issues which require further exploration. In terms of classification, there is the need to design innovative algorithms for the early detection of data anomalies (outliers, concept drifts and missing data) and optimise the data partition sizes.

Regarding recommendation, it is necessary to develop different algorithms and new approaches because most of the contributions for data streams are based on matrix factorisation. The problem of hyper-parameter optimisation, which is common to classification and recommendation, has a large impact on the performance of the algorithms and, thus, needs to be further addressed. Currently, researchers are giving the initial steps on progressive automation of machine learning (AutoML) in order to substitute machine learning experts, including automated hyper-parameter tuning. In terms of future developments, there are two critical points requiring further research: the detection user preference evolution and the scalability of the on-line algorithms. While the first will lead to the design of more robust and efficient forgetting techniques, the latter will rely on new model compressing techniques to reduce the data dimension and on the optimisation of the parallel implementation of recommendation engines.

Finally, classification and recommendation algorithms need to embrace further the lack of information problem, including both the non-arrival and the late arrival of events, as well as the design continuous self-tuning incremental models, *e.g.*, hyper-parameter meta-learners.

In terms of evaluation of the recommendation with data streams, it is necessary develop evaluation mechanisms not only focussed on performance parameters like predictive and accuracy metrics, but also on the algorithm runtime and computational resources used, to find the necessary commitment between functional and non-functional parameters. Additionally, it is important to evaluate the learning behaviour of the algorithms.

CONCLUSION

The presented techniques are intended for scenarios where the information is dynamic, continuous and unknown. Their main goal is to build and maintain updated models, capable of the capturing the relevant patterns as they occur, to make accurate predictions. The foundations of data streams algorithms are incremental processing, to reduce the computational requirements required to process large quantities of data, and relevance model updating, adjust the data stream model to the current user trends. In terms of research, this topic covers a large set of real problems, including sensor networks, crowdsourced data and web analysis. The numerous fields of application illustrate the relevance of stream-based knowledge techniques, ranging from industrial sensor networks to financial and e-commerce platforms.

Recently, there are several alternatives proposed by researchers for three core topics of data stream processing: classification, recommendation and evaluation. The tendency is to create more advance and accurate algorithms with the help of increased computational resources, *e.g.*, cloud computing. Data anomalies, *e.g.*, concept drifts, outliers and missing data, contribute to the decrease of the overall performance of the algorithms, suggesting the need to design novel and more advanced techniques, supported by automated learning, to dynamically adapt to the behaviour of the incoming data. Specifically, we anticipate that computer processing power, parsimonious data representation and innovative self-tuning algorithms will be essential for the near future of data stream processing. Finally, the evalu-

ation of stream processing techniques requires further exploration since the available techniques are extensions of off-line evaluation.

REFERENCES

- Aggarwal, C. C. (Ed.). (2007). *Data streams: models and algorithms* (Vol. 31). Springer Science & Business Media. doi:10.1007/978-0-387-47534-9
- Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S. (2004). On demand classification of data streams. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 503-508). ACM.
- Angiulli, F., & Fasseti, F. (2007). Detecting distance-based outliers in streams of data. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 811-820). ACM. 10.1145/1321440.1321552
- Anyosa, S. C., Vinagre, J., & Jorge, A. M. (2018). Incremental Matrix Co-factorization for Recommender Systems with Implicit Feedback. In *Companion of the The Web Conference 2018 on The Web Conference 2018* (pp. 1413-1418). International World Wide Web Conferences Steering Committee. 10.1145/3184558.3191585
- Bifet, A. (2010). Adaptive stream mining: Pattern learning and mining from evolving data streams. In *Proceedings of the 2010 conference on adaptive stream mining: Pattern learning and mining from evolving data streams* (pp. 1-212). IOS Press.
- Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning* (pp. 89-96). ACM.
- Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 39-46). ACM. 10.1145/1864708.1864721
- Datar, M., Gionis, A., Indyk, P., & Motwani, R. (2002). Maintaining stream statistics over sliding windows. *SIAM Journal on Computing*, 31(6), 1794–1813. doi:10.1137/S0097539701398363
- Duarte, J., Gama, J., & Bifet, A. (2016). Adaptive model rules from high-speed data streams. *ACM Transactions on Knowledge Discovery from Data*, 10(3), 30. doi:10.1145/2829955
- Galić, Z. (2016). *Spatio-temporal data streams*. Springer. doi:10.1007/978-1-4939-6575-5
- Gama, J., & Gaber, M. M. (Eds.). (2007). *Learning from data streams: processing techniques in sensor networks*. Springer Science & Business Media. doi:10.1007/3-540-73679-4
- Gama, J., Sebastião, R., & Rodrigues, P. P. (2009). Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 329-338). ACM. 10.1145/1557019.1557060
- Garofalakis, M., Gehrke, J., & Rastogi, R. (Eds.). (2016). *Data Stream Management: Processing High-Speed Data Streams*. Springer. doi:10.1007/978-3-540-28608-0

- Ghanem, T. M., Hammad, M. A., Mokbel, M. F., Aref, W. G., & Elmagarmid, A. K. (2007). Incremental evaluation of sliding-window queries over data streams. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 57–72. doi:10.1109/TKDE.2007.250585
- Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfahringer, B., Holmes, G., & Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. *Machine Learning*, 106(9-10), 1469–1495. doi:10.1007/10994-017-5642-8
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. doi:10.1148/radiology.143.1.7063747 PMID:7063747
- Hopfgartner, F., Kille, B., Heintz, T., & Turrin, R. (2015). Real-time recommendation of streamed data. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 361-362). ACM.
- Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97-106). ACM. 10.1145/502512.502529
- Jaworski, M., Duda, P., Rutkowski, L., Najgebauer, P., & Pawlak, M. (2017). Heuristic regression function estimation methods for data streams with concept drift. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 726-737). Springer. 10.1007/978-3-319-59060-8_65
- Klinkenberg, R. (2001). Using labeled and unlabeled data to learn drifting concepts. In *Workshop notes of the IJCAI-01 Workshop on Learning from Temporal and Spatial Data* (pp. 16-24). Academic Press.
- Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4), 89–97. doi:10.1145/1721654.1721677
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30–37. doi:10.1109/MC.2009.263
- Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37, 132–156. doi:10.1016/j.inffus.2017.02.004
- Li, J., Maier, D., Tufte, K., Papadimos, V., & Tucker, P. A. (2005). No pane, no gain: Efficient evaluation of sliding-window aggregates over data streams. *SIGMOD Record*, 34(1), 39–44. doi:10.1145/1058150.1058158
- Matuszyk, P., & Spiliopoulou, M. (2017). Stream-based semi-supervised learning for recommender systems. *Machine Learning*, 106(6), 771–798. doi:10.1007/10994-016-5614-4
- Matuszyk, P., Vinagre, J., Spiliopoulou, M., Jorge, A. M., & Gama, J. (2017). Forgetting techniques for stream-based matrix factorization in recommender systems. *Knowledge and Information Systems*, 1–30.
- Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., & Owen, S. (2016). Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17(1), 1235–1241.
- Montiel, J., Read, J., Bifet, A., & Abdessalem, T. (2018). Scalable Model-Based Cascaded Imputation of Missing Data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 64-76). Springer. 10.1007/978-3-319-93040-4_6

- Muthukrishnan, S. (2005). Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2), 117-236.
- Osojnik, A., Panov, P., & Džeroski, S. (2017). Multi-label classification via multi-target regression on data streams. *Machine Learning*, 106(6), 745–770. doi:10.1007/10994-016-5613-5
- Paterek, A. (2007). Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop* (Vol. 2007, pp. 5-8). Academic Press.
- Pfaffe, P., Tillmann, M., Walter, S., & Tichy, W. F. (2017). Online-Autotuning in the Presence of Algorithmic Choice. In *Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017 IEEE International* (pp. 1379-1388). IEEE. 10.1109/IPDPSW.2017.28
- Pokrajac, D., Lazarevic, A., & Latecki, L. J. (2007). Incremental local outlier detection for data streams. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on* (pp. 504-515). IEEE. 10.1109/CIDM.2007.368917
- Rutkowski, L., Pietruczuk, L., Duda, P., & Jaworski, M. (2013). Decision trees for mining data streams based on the McDiarmid's bound. *IEEE Transactions on Knowledge and Data Engineering*, 25(6), 1272–1279. doi:10.1109/TKDE.2012.66
- Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8(Jul), 1623–1657.
- Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. *Journal of Machine Learning Research*, 10(Mar), 623–656.
- Veloso, B., Malheiro, B., & Gama, J. (2018). Self Hyper-Parameter Tuning for Data Streams. 21st International Conference on Discovery Science (DS 2018). *Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence*.
- Vinagre, J., Jorge, A. M., & Gama, J. (2014). Online bagging for recommender systems. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*. Advance online publication. doi:10.1111/exsy.12303
- Vinagre, J., Jorge, A. M., & Gama, J. (2018). Fast incremental matrix factorization for recommendation with positive-only feedback. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 459-470). Springer.
- Wang, H., Fan, W., Yu, P. S., & Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 226-235). ACM. 10.1145/956750.956778
- Wang, W., Yin, H., Huang, Z., Wang, Q., Du, X., & Nguyen, Q. V. H. (2018). Streaming Ranking Based Recommender Systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 525-534). ACM. 10.1145/3209978.3210016

ADDITIONAL READING

Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavalda, R. (2009). New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 139-148). ACM. 10.1145/1557019.1557041

Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 71-80). ACM. 10.1145/347090.347107

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: A review. *SIGMOD Record*, 34(2), 18–26. doi:10.1145/1083784.1083789

Gama, J. (2010). *Knowledge discovery from data streams*. CRC Press. doi:10.1201/EBK1439826119

Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. In *Brazilian symposium on artificial intelligence* (pp. 286-295). Springer, Berlin, Heidelberg.

Gama, J., Sebastião, R., & Rodrigues, P. P. (2009). Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 329-338). ACM. 10.1145/1557019.1557060

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9), 2250–2267. doi:10.1109/TKDE.2013.184

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97-106). ACM. 10.1145/502512.502529

Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 426-434). ACM. 10.1145/1401890.1401944

Muthukrishnan, S. (2005). Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2), 117-236.

KEY TERMS AND DEFINITIONS

Automatic Machine Learning: Is the automation of complex data mining tasks for example hyper-parameter optimization.

Classification: Is the task that assigns items to a set of classes.

Concept Drift: Represents the change over time of the distribution of a target variable.

Data Mining: Is the process that identifies and collects patterns in data sets.

Data Streams: is a continuous flow of data to be processed by an algorithm.

Evaluation: Is the task of judge and assessing if the predictions generated by a classification or recommendation algorithm scores on the target value.

Hyper-Parameter: It is used to describe the parameters that an algorithm/model requires to work.

Recommendation: Is an algorithm that tries to identify patterns and make predictions based on the past user behaviour.

N-Tuple Algebra as a Generalized Theory of Relations

Boris A. Kulik

Institute for Problems in Mechanical Engineering of RAS, Russia

Alexander Y. Fridman

 <https://orcid.org/0000-0003-2408-6892>

Institute for Informatics and Mathematical Modeling of RAS, Russia

INTRODUCTION

In information technologies, analysis of heterogeneous information often necessitates unification of presentation forms and processing procedures for such data. One of the promising ways to solve this problem is the search for a universal structure, which allows reducing various data and knowledge models to a single mathematical model with unified analysis methods. One of such universal structures is the relation, which is mainly associated with databases modeled by means of relational algebra (Codd, 1971). However, relations are used not only in databases. They can model as different, at first glance, mathematical objects as graphs, networks, artificial intelligence structures, predicates, logical formulas, etc. Representation and analysis of such structures and models requires for more expressive means and methods than relational algebra provides.

So, with a view to developing a general theory of relations, the authors propose n -tuple algebra (NTA) developed as a theoretical generalization of structures and methods applicable in intelligence systems. NTA allows for formalizing a wide set of logical problems (deductive, abductive and modified reasoning, modeling uncertainties and so on). Besides, this paper considers matters of metrization and clustering for NTA objects with ordered domains of attributes.

BACKGROUND

Nowadays, intelligence systems include two types of dissimilar objects, namely databases (DBs) and knowledge bases (KBs). Their structures are fundamentally different, since their construction is based on different theoretical approaches. Modern data structures (numbers, vector graphs, networks, texts, etc.) exploit algebraic techniques (Wirth, 1976). As for KBs, their basic models (predicates, frames, semantic networks, rules) are built on the basis of declarative approaches (Russel & Norvig, 2003). This discrepancy results in significant differences of programming systems for DBs and KBs and, accordingly, in large expenditures of time and money to couple DBs and KBs in one software system. Conventional cases to attempt this integration are mostly accomplished within declarative approaches. An example is the deductive database model (Ullman, 1989; Ceri, Gotlob, & Tanca, 1990), which is based on a modified Prolog programming language intended for computer implementation of the declarative approach.

Conversely, some shortcomings inhere in this approach. One of the main problems is that when using a declarative approach, many tasks of logical analysis need to be reduced to satisfiability checks for a

DOI: 10.4018/978-1-7998-3479-3.ch048

certain logical formula where only two possible answers (“yes” or “no”) are possible. Such a reduction is not simple. Moreover, it is unrealizable in cases when one needs to not only receive a binary answer but also to estimate the value of some parameters in the formal system or to assess the structure and/or number of objects that satisfy the given conditions.

When modeling and analyzing modifiable reasoning (i.e., reasoning with hypotheses and abductive conclusions), many researchers recommend using non-classical logic (default logic, non-monotonic logic, etc.) (Russel & Norvig, 2003), since it is believed that classical logic is ill-suited for solving such problems. However, in many cases, interpretations of non-classical logics are either missing or not corresponding to the semantics of the objects being modeled. There are reasons to assume that the difficulties of modeling modifiable reasoning within the framework of classical logic are not due to the shortcomings of this logic, but to the specifics of the declarative approach.

Algebraic approach to logical modeling can be originated from Aristotle’s syllogistics and Boole’s logical matrices, in more recent times only Arkady Zakrevskij (for instance, see (Zakrevskij & Gavrilov, 1969)) investigated set-theoretical approach to logic and proposed a programming language LYaPAS (Logical Language for Representation of Synthesis Algorithms) oriented toward programming of synthesis algorithms for finite-state and discrete devices. However, this language does not provide any means for logical inference.

To overcome the mentioned and other problems, the mathematical model of relational algebra was generalized by the authors to a more universal system for processing n -ary relations called n -tuple algebra (Kulik, 1995, 2007; Kulik & Fridman, 2017). Its main features and provided opportunities are introduced below.

N-TUPLE ALGEBRA: BASICS AND FEATURES

The algebraic approach to modeling systems stipulates that some homogeneous objects (numbers, vectors, matrices, etc.) are used as basic structures. In n -tuple algebra (NTA), such objects are relations, which can be expressed as four types of structures called **NTA objects**. Every NTA object is immersed into a certain space of variables called **attributes**. Each attribute is represented by a set of possible values called **domains**, and the spaces themselves are defined as Cartesian products of these domains.

Names of NTA objects contain an identifier followed by a sequence of attributes names in square brackets; these attributes determine the **relation diagram** in which the NTA object is defined. For example, $R[XYZ]$ denotes an NTA object defined within the space of attributes $X \times Y \times Z$ that is $R[XYZ] \in X \times Y \times Z$.

From the mathematical point of view, NTA is an **algebraic system** $A = \langle A, O, R \rangle$, where the **carrier** A is an arbitrary totality of relations expressed as NTA objects; O is a set of **operations** comprising **algebra of sets’ operations** (union, intersection, complement) and **operations on attributes** (see below); R is a set of verified **relations** between NTA objects, namely equality and inclusion. This algebraic system is proved to be complete (i.e. all operations and checks of relations are feasible for any totalities of NTA objects).

NTA objects defined on the same relation diagram are called **homotypic** ones. NTA objects contain n -tuples of sets and provide a concise representation of n -ary relations. In NTA, **elementary n -tuples** correspond to n -tuples of elements in ordinary relations. For instance, the record $T[XYZ] = (a, b, c)$ means that T is an elementary n -tuple where $a \in X, b \in Y, c \in Z$.

NTA objects look like matrices having subsets of corresponding attributes (**components**) rather than elements as their cells. The totality of NTA components includes two **dummy components**. They are: the **complete component** “*” that equals to the whole domain (i.e. the range of values) of the corresponding attribute in the relation diagram and the **empty component** “ $\emptyset$ ” that equals to the empty set and models the complement of the component “*”.

NTA admits four types of NTA objects, namely C -n-tuples and C -systems, D -n-tuples and D -systems.

A **C -n-tuple** is an n -tuple of components defined in a certain relation diagram, denoted within square brackets and interpreted as the set of elementary n -tuples contained in the Cartesian product of its components. For example, the right part of the formula $R[XYZ] = [A * C]$ expresses a C -n-tuple that can be transformed into an ordinary n -ary relation by calculating the Cartesian product $A \times * \times C$ where $A \subseteq X$, $* = Y$ is a complete component placed to the position of the attribute Y , and $C \subseteq Z$.

A **C -system** is a set of homotypic C -n-tuples that are denoted as a matrix in square brackets. As an example, the C -system $Q[XYZ] = \begin{bmatrix} A_1 & B_1 & * \\ * & B_2 & C_2 \end{bmatrix}$ can be reduced to a set of elementary n -tuples calculated as $Q[XYZ] = (A_1 \times B_1 \times Z) \cup (X \times B_2 \times C_2)$.

A **D -n-tuple** expresses a structure that can be interpreted, in particular, as a complement of a certain C -n-tuple. Suppose, the C -n-tuple $R[XYZ] = [A B C]$ is given. Its complement equals to the difference of the C -n-tuples $[X Y Z] \setminus [A B C] = (X \times Y \times Z) \setminus (A \times B \times C)$. The formula of difference for two Cartesian products is well known, for the given example it equals:

$$(X \times Y \times Z) \setminus (A \times B \times C) = ((X \setminus A) \times Y \times Z) \cup (X \times (Y \setminus B) \times Z) \cup (X \times Y \times (Z \setminus C)).$$

In NTA terms, the right part of this formula can be denoted as the C -system

$$\bar{R}[XYZ] = \begin{bmatrix} \bar{A} & * & * \\ * & \bar{B} & * \\ * & * & \bar{C} \end{bmatrix}$$

where $\bar{A}$, $\bar{B}$, $\bar{C}$ mean complements of the components A , B , C to the domains of corresponding attributes.

The square C -systems like $\bar{R}[XYZ]$ where all nondiagonal cells contain complete components are called **diagonal C -systems**. Without any loss of information, they can be written as an n -tuple of the diagonal components. To distinguish them from C -n-tuples in NTA, we use reversed square brackets instead of regular ones. So, for the given example $\bar{R}[XYZ] =] \bar{A} \bar{B} \bar{C} [$.

A **D -system** is a structure that consists of a set of homotypic D -n-tuples and equals to their intersection. The list of D -n-tuples in a D -system is framed with reversed square brackets.

In order to better understand relationships between different types of NTA objects, consider an example. Let the attribute space $[XYZ]$ be given, where $X = \{a, b, c, d\}$, $Y = \{f, g, h\}$, $Z = \{a, b, c\}$. There is a C -system defined in this space:

$$F[XYZ] = \begin{bmatrix} \{a, b, d\} & \{f, h\} & \{b\} \\ \{b, c\} & * & \{a, c\} \end{bmatrix}.$$

Then the complement of $F[XYZ]$ is the D -system

$$\bar{F}[XYZ] = \begin{bmatrix} X \setminus \{a, b, d\} & Y \setminus \{f, h\} & Z \setminus \{b\} \\ X \setminus \{b, c\} & Y * & Z \setminus \{a, c\} \end{bmatrix} = \begin{bmatrix} \{c\} & \{g\} & \{a, c\} \\ \{a, d\} & \emptyset & \{b\} \end{bmatrix}.$$

In turn, according to the duality law in algebra of sets, since the C -system $F[XYZ]$ is the union of the C - n -tuples contained in it, the D -system $\bar{F}[XYZ]$ can be represented as the intersection of the D - n -tuples it contains:

$$\bar{F}[XYZ] =]\{c\} \{g\} \{a, c\}[\cap]\{a, d\} \emptyset \{b\}[.$$

NTA laws are based on the laws of algebra of sets and properties of the Cartesian product. Let us introduce main NTA relationships.

Suppose two homotypic C - n -tuples $P = [P_1 P_2 \dots P_n]$ and $Q = [Q_1 Q_2 \dots Q_n]$ be given. Then:

- 1) $P \subseteq Q$, if and only if $P_i \subseteq Q_i$ for all $i = 1, 2, \dots, n$;
- 2) $P \cap Q = [P_1 \cap Q_1 P_2 \cap Q_2 \dots P_n \cap Q_n]$. Moreover, if for any i $P_i \cap Q_i = \emptyset$, then $P \cap Q = \emptyset$;
- 3) $P \cup Q = \begin{bmatrix} P_1 & P_2 & \dots & P_n \\ Q_1 & Q_2 & \dots & Q_n \end{bmatrix} \subseteq [P_1 \cup Q_1 P_2 \cup Q_2 \dots P_n \cup Q_n]$;
- 4) equality $P \cup Q = [P_1 \cup Q_1 P_2 \cup Q_2 \dots P_n \cup Q_n]$ is possible in two cases only:
 - (i) $P \subseteq Q$ or $Q \subseteq P$;
 - (ii) $P_i = Q_i$ for all corresponding pairs of components except one pair;
- 5) $\bar{P} =]\bar{P}_1 \bar{P}_2 \dots \bar{P}_n[.$

Numerous examples of representing different data and knowledge structures as NTA objects are given in the references to our publications listed below. Due to size limits of this paper, we give only one example of how to apply NTA structures for logical analysis. This example uses variables that contain more than two values, which makes it possible to analyze the logic models that go beyond the propositional calculus. Suppose we want to find the object, which is in one of three rooms. At the same time, you may not enter one of the rooms (there is a danger), and one of the rooms is empty. There are two following clues.

- 1) The first room is not dangerous, the second is not empty;
- 2) The third room is not dangerous, the second is not empty.

We also know that one of the clues is true, the other one is false.

Let us denote A the statement “the room contains the desired object”, B – “the room is empty” and C – “the room is dangerous”. Attributes names will be equal to the names of the rooms R_1, R_2, R_3 . Each of these rooms corresponds to one of situations from the permitted situations set $\{A, B, C\}$.

The given conditions can be expressed by C - n -tuples as follows.

- 1) $T_1[R_1 R_2 R_3] = [\{A, B\} \{A, C\} *];$
- 2) $T_2[R_1 R_2 R_3] = [* \{A, C\} \{A, B\}].$

To solve the problem, it is necessary to consider two possible cases: 1) The first clue is true, and the second is false; 2) The first clue is false, and the second is true.

Let us consider the first case. If T_2 is false, its negation $\bar{T}_2 =] \emptyset \{B\} \{C}[$ is true. To find a solution, one needs to find the intersection of the true conditions:

$$T_1 \cap \bar{T}_2 =]\{A, B\} \{A, C\} *] \cap] \emptyset \{B\} \{C}[=]\{A, B\} \{A, C\} *] \cap \begin{bmatrix} * & \{B\} & * \\ * & * & \{C\} \end{bmatrix} =]\{A, B\} \{A, C\} \{C}[.$$

The resulting C -n-tuple contains 4 possible placement options (one can find this by calculating the Cartesian product of $\{A, B\} \times \{A, C\} \times \{C\}$). However, according to the conditions of the problem, the correct placement must have different situations in different rooms, i.e. the only placement (B, A, C) is correct. It means that the desired object is located in the second room. Similar results can be obtained when testing the second case.

The above listed relationships provide proving all the rest NTA laws, which assisted in composing algorithms of NTA operations and checks of equality and inclusion for any homotypic NTA objects.

To implement NTA operations and checks for any (possibly non-homotypic) NTA objects, it is necessary to define generalized operations and relations based on some operations upon attributes.

NTA: GENERALIZED OPERATIONS AND OPERATIONS UPON ATTRIBUTES

Let us introduce operations on attributes first.

Renaming of attributes is used when it is necessary to substitute variables, particularly, in algorithms for calculating transitive closure of a graph.

Transposition of attributes is an operation that swaps columns in an NTA object's matrix and respectively changes the order of attributes in the relation diagram.

If an NTA object contains two columns, the permutation of attributes without permutation of columns (or, equivalently, permutation of columns without permutation of attributes) corresponds to the operation of **reversing relations**.

Elimination of an attribute ($-Attr$) is done in the following way: a column is removed from an NTA object, and the corresponding attribute is removed from the relation diagram.

Addition of a dummy attribute ($+Attr$) is done when the added attribute is not yet present in the relation diagram of an NTA object. This operation adds the name of a new attribute into the relation diagram and simultaneously adds a new column with dummy components into the corresponding place; dummy components “*” are added into C -n-tuples and C -systems, and dummy components “ $\emptyset$ ” are added into D -n-tuples and D -systems.

By defining the operation $+Attr$, we have eliminated the restriction in the theory of relations stating that algebra-of-sets laws are only applicable to the relations defined upon the same Cartesian product.

The operation $+Attr$ is often used to reduce some different-type NTA objects to the same relation diagram by adding the missed attributes. Then one can perform all necessary operations and checks by means of standard NTA algorithms.

Considering this, we have introduced **generalized operations** ($\cap_G, \cup_G$). They differ from corresponding operations of algebra of sets by the only feature: NTA objects are reduced to the same relation diagram

before execution of these operations, which are semantically equal to logical connectives: disjunction and conjunction correspondingly. **Generalized relations** ($\in_G$ and $=_G$) are defined similarly.

Algebra of n -ary relations with the mentioned generalized operations and relations is proved to be isomorphic to the conventional algebra of sets.

NTA: CORRESPONDENCE WITH PROPOSITIONAL AND PREDICATE CALCULI

A C - n -tuple corresponds to the **conjunction** of unary predicates with different variables. For example, in predicate calculus, the C - n -tuple $P[XYZ] = [P_1 P_2 P_3]$ where $P_1 \subseteq X$; $P_2 \subseteq Y$; $P_3 \subseteq Z$ models the logical formula $H = P_1(x) \wedge P_2(y) \wedge P_3(z)$ where X , Y and Z are ranges of values for the variables x , y and z .

A D - n -tuple corresponds to the **disjunction** of unary predicates with different variables. For instance, the D - n -tuple $Q[XYZ] =]Q_1 Q_2 Q_3[$ corresponds to the logical formula

$$F_2(x, y, z) = Q_1(x) \vee Q_2(y) \vee Q_3(z).$$

C -systems are convenient for representing **disjunctive normal forms** (DNFs) of unary predicates, and D -systems model **conjunctive normal forms** (CNFs) of unary predicates.

An elementary n -tuple that is a part of a non-empty NTA object corresponds to a **satisfying substitution** in a logical formula.

An empty NTA object corresponds to an **identically false formula**.

An NTA object $P[XY...Z]$ that equals to $X \times Y \times \dots \times Z$ corresponds to a **valid formula**, or a **tautology**.

A non-empty NTA object models a **satisfiable formula**.

Now let us consider quantifiers in NTA.

Adding a dummy attribute to a C - n -tuple or a C -system corresponds to the **rule of generalization** in predicate calculus. For example, if an NTA object

$$G[XZ] = \begin{bmatrix} \{a, c\} & * \\ \{a, c, d\} & \{b, c\} \end{bmatrix}$$

corresponds to a formula $F(x, z)$ of this calculus, by adding a dummy attribute Y into this NTA object we get an NTA object

$$G_1[XYZ] = +Y(G[XZ]) = \begin{bmatrix} \{a, c\} & * & * \\ \{a, c, d\} & * & \{b, c\} \end{bmatrix}$$

which models the formula $\forall y F(x, z)$.

Elimination of an attribute from a C -system (D -system) is proved to be equivalent to the universal (existential) quantification on this attribute. Thus, if an NTA object $R[XYZ]$ models a formula $F(x, y, z)$, then $-Y(R[XYZ])$ corresponds to $\exists y F(x, y, z)$, if $R[XYZ]$ is a C -system, and to $\forall y F(x, y, z)$, if $R[XYZ]$ is a D -system.

To obtain NTA representations for formulas of propositional calculus containing logical variables $x_1, x_2, \dots, x_n$, it is enough to assign every variable x_i with the attribute $X_i = \{0, 1\}$. Then the dummy com-

ponent * corresponds to the set $\{0, 1\}$, literal x_i corresponds to $\{1\}$, and its negation $\neg x_i$ corresponds to $\{0\}$. For example, the formula

$$F(x, y, z) = (x \vee \neg z) \wedge (\neg y \vee z) \text{ corresponds to the } D\text{-system } R[XYZ] = \begin{bmatrix} \{1\} & \emptyset & \{0\} \\ \emptyset & \{1\} & \{0\} \end{bmatrix}.$$

The negation of this formula

$$\neg F(x, y, z) = (\neg x \wedge z) \vee (y \wedge \neg z) \text{ is modeled by the } C\text{-system } \bar{R}[XYZ] = \begin{bmatrix} \{0\} & * & \{1\} \\ * & \{0\} & \{1\} \end{bmatrix}.$$

It is easy to see that relations between C -objects (C - n -tuples and C -systems) and D -objects (D - n -tuples and D -systems) are in accordance with de Morgan's **laws of duality**. Due to this fact, they are called **alternative classes**.

All NTA objects model scopes of truth for their corresponding logical formulas. In NTA, attribute domains can contain arbitrary sets of values that are not necessarily equal to each other. Thus, NTA structures model formulas of **many-sorted predicate calculus**.

This paper does not consider matters of complexity in detail; they are described in (Kulik, 1995). Here, we would like to note that NTA structures can be polynomially reduced to logical ones; hence *the computational complexity of algorithms on NTA structures fully corresponds to the computational complexity of algorithms solving problems on logical structures*. A significant number of such problems arising during logical analysis by means of deduction procedures (for instance, the satisfiability problem for a CNF) are NP-complete problems with regard to their computational complexity. However, the matrix properties of NTA objects allowed us to find some special cases that are solvable in polynomial time only (Kulik, 1995). Moreover, such structure is good to reduce computational complexity of NTA operations by their effective parallelization (Kulik & Fridman, 2014).

LOGICAL INFERENCE IN NTA

Theoretical principles of NTA provide solving the following problems of deductive analysis.

- 1) *Problem of correctness check for an alleged consequence B from the given premises A_i .*
- 2) *Problem of deriving possible consequences from the given premises A_i considering certain semantical constraints, for instance, presence of given variables or their combinations in a consequence, deriving a consequence with the minimal number of significant variables, etc.*

Unlike other logical systems that solve such problems by using inference rules with hardly optimized order of their implementation, NTA solves these problems by means of certain standard algorithms. To do so, we transform classical logic' formulas, which express premises and consequences, into NTA objects and subject these objects to generalized operations and checks of equality ($=_G$) or inclusion ($\in_G$). The transition to the algebraic representation becomes clearer, if you consider that NTA objects model scopes of truth for logical formulas. Then the correctness proof for an alleged consequence B from the premises A_i requires for calculating generalized intersections and checking the following generalized inclusion:

$$(A_1 \cap_G \dots \cap_G A_n) \subseteq_G B \quad (1)$$

Considering (1), the problem to derive possible consequences $\{B_j\}$ from the given premises A_i can be solved in NTA by using the relation that must be true for every B_j : $A \subseteq B_j$, where $A = A_1 \cap_G \dots \cap_G A_n$. Thus, any correct consequence from premises A_i can be modeled by an NTA object representing a superset of A . To find such objects, NTA proposes several techniques. One of them is **calculating projections of an NTA object**. This operation is simple if A is a C -system. Then the projection can be found by operation $-Attr$. It is easy to prove that any projection of an NTA object models its superset. This technique allows checking the possibility to obtain consequences containing certain variables.

The relationship (1) allowed for an NTA definition of the **minimal consequence** from axioms $A_1, A_2, \dots, A_n$ that equals to the NTA object $A = A_1 \cap_G A_2 \cap_G \dots \cap_G A_n$. This consequence is the minimal one as it is proved to satisfy the two following statements:

- 1) any strict subset of A is not derivable from the axioms $A_1, A_2, \dots, A_n$;
- 2) if S is an arbitrary NTA object from a proper universe, the NTA object $A \cup S$ is derivable from the axioms $A_1, A_2, \dots, A_n$.

In some kinds of reasoning models, a consequence B is considered a suitable solution of an inference task if this consequence contains fewer variables than the minimal consequence A (such a consequence can be found by using the above-introduced technique. This decrease in number of variables in the formula $B = A \cup S$ results from transformation of some variables in the formula A into dummy ones after adding S to it.

NTA: MODELS OF PLAUSIBLE REASONING

Some difficulties in applying propositional and predicate calculi to analysis of plausible reasoning arose from the fact that mathematical logic recognizes inconsistency of a reasoning system (a theory) if and only if both a corollary and its negation follow from the same premises. Conversely, commonsense reasoning recognize inference of contrary corollaries (for instance, both statements “All A are B ” and “No A are B ”) as a definite criterion of inconsistency for a knowledge system though they are not formally contradictory.

In order to eliminate this and other discrepancies between formal logic and natural deduction, we propose a concept of “**collision**” in our logical analysis system (Kulik, 1995).

Unlike a logical contradiction, collisions can have opposite interpretations in different cases. For example, within one system the equality $R = \emptyset$ means the absence of an object that is necessary for existence of the system, and in another system (where the object R is not necessary) this equality specifies a status of this object. The first case requires for changing the premises while the second case provides a new useful knowledge.

Collisions mostly occur during defeasible reasoning when new knowledge or a hypothesis is included into the logical system. Collisions indicate violations of some formal rules or restrictions providing consistency of this system.

By using collisions, NTA allows for a formal definition of hypotheses. Let us suppose that a system of premises expressed as NTA objects $A_1, A_2, \dots, A_n$ is given and the NTA object $A = A_1 \cap_G A_2 \cap_G \dots \cap_G A_n$ is calculated.

Then a certain formula H is called a ***hypothesis*** under the following conditions:

- 1) $A \subseteq_G H$ is false. Otherwise, H is a consequence according to (1).
- 2) The hypothesis is correct, if the object $H \cap_G A$ contains no collisions.

The first condition can be expressed in other way: $A \setminus_G H \neq \emptyset$. The second condition assumes that the hypothesis is a premise or an axiom.

Usually, forming and checking of hypotheses accompany other analysis methods for defeasible reasoning. Below, we will describe usage of hypotheses in searching for abductive conclusions.

Abduction is forming of an explanatory hypothesis when one knows some of premises and an estimated consequence confirmed with facts or reasonable arguments, but a formal check does not infer it from the given premises. For example, abduction is used during diagnostics.

Let us now formally define abduction.

If B is an estimated consequence from the premises $A = A_1 \cap_G A_2 \cap_G \dots \cap_G A_n$ and the statement $A \subseteq_G B$ is known to be false, then a formula H is an ***admissible abductive conclusion*** when the two following conditions are met:

- 1) H is a hypothesis (i.e. $A \subseteq_G H$ is false) and $H \cap_G A$ is not empty;
- 2) $(H \cap_G A) \subseteq_G B$, that is adding H into the system of premises results in deducibility of the estimated consequence B .

Considering the above definitions, we propose a search algorithm for abductive conclusions.

Consider an example. The goal is to check the correctness of inference for the following argument: “If Jones did not meet Smith that night, than either Smith was a killer, or Jones is lying. If Smith was not the killer, then Jones did not meet Smith this night and the murder happened after midnight. If the murder happened after midnight, then either Smith was the killer, or Jones is lying. Consequently, Smith was the killer”.

First, we express this argument in the language of propositional calculus. Introduce the following notation: C , Jones met Smith that night; D , Smith was the murderer; E , Jones is lying; F , the murder happened after midnight. Then, after converting each of the proposals into a DNF, we obtain the following expressions:

- for the first sentence: $C \vee (D \wedge \neg E) \vee (\neg D \wedge C)$;
- for the second sentence: $D \vee (\neg C \wedge F)$;
- for the third sentence: $\neg F \vee (D \wedge \neg E) \vee (\neg D \wedge E)$;
- for the conclusion: D .

In order to rewrite these formulas as NTA objects, we use the universe $X_C \times X_C \times X_E \times X_F = \{0, 1\}^4$ where $C = D = E = F = 1$ and $\neg C = \neg D = \neg E = \neg F = 0$. Then, the premises are expressed as the following C -systems:

$$A_1 = \begin{bmatrix} \{1\} & * & * & * \\ * & \{1\} & \{0\} & * \\ * & \{0\} & \{1\} & * \end{bmatrix}; A_2 = \begin{bmatrix} * & \{1\} & * & * \\ \{0\} & * & * & \{1\} \end{bmatrix}; A_3 = \begin{bmatrix} * & * & * & \{0\} \\ * & \{1\} & \{0\} & * \\ * & \{0\} & \{1\} & * \end{bmatrix}$$

and the conclusion is expressed as the C -n-tuple $B[X_D] = [\{1\}]$.

To solve the problem, we check the relation $(A_1 \cap_G A_2 \cap_G A_3) \subseteq_G B[X_D]$. To do this, we calculate

$$A[X_C X_D X_E X_F] = A_1 \cap_G A_2 \cap_G A_3 = \begin{bmatrix} \{1\} & \{1\} & * & \{0\} \\ * & \{1\} & \{0\} & * \\ \{0\} & \{0\} & \{1\} & \{1\} \end{bmatrix}.$$

To check the inclusion, we add the missing attributes in B : $B[X_C X_D X_E X_F] = [* \{1\} * *]$.

Verification shows that the third C -n-tuple of A is not included in B , i.e., the supposed conclusion (Smith was the killer) is not deducible. In order to confirm or refute the accuracy of the conclusion, certain circumstances must be clarified. The search for such circumstances can be formulated as the problem of finding an abductive conclusion.

Suppose that the conclusion is correct. Then, we need to find hypotheses whose addition to the premises will result in deducibility of the supposed conclusion.

To find such hypotheses, we use the algorithm proposed above:

$$R = A \setminus_G B = \begin{bmatrix} \{1\} & \{1\} & * & \{0\} \\ * & \{1\} & \{0\} & * \\ \{0\} & \{0\} & \{1\} & \{1\} \end{bmatrix} \cap_G [* \{0\} * *] = [\{0\} \{0\} \{1\} \{1\}].$$

Now, we can select any projection of R as R_i . Let it be $R[X_F]$, then $R_i = [* * * \{1\}]$, $H_i = [* * * \{0\}]$.

Since no additional restrictions are defined for this example, it is sufficient to check whether the intersection of premises becomes confluent for the chosen hypothesis:

$$A \cap_G H_i = \begin{bmatrix} \{1\} & \{1\} & * & \{0\} \\ * & \{1\} & \{0\} & * \\ \{0\} & \{0\} & \{1\} & \{1\} \end{bmatrix} \cap_G [* * * \{0\}] = \begin{bmatrix} \{1\} & \{1\} & * & \{0\} \\ * & \{1\} & \{0\} & \{0\} \end{bmatrix}.$$

Verification confirms the correctness of this hypothesis. In natural language, the tested hypothesis means that the murder happened before midnight. Hence, the conclusion would be correct if a clarification of the murder time would comply with the hypothesis.

Often, it is not sufficient to test one hypothesis for analysis of an argument. For the above example in particular, another possible hypothesis arises from selecting the projection $[X_C X_E]$. Then the conclusion would be correct, if the following hypothesis is confirmed: “Jones met Smith that night and Jones does not lie”.

MEASURES IN NTA

NTA structures consist of sets of n -tuples being Cartesian products in essence. A measure (a probabilistic measure in particular) can be easily calculated for a Cartesian product; that is why NTA structures can

be readily immersed into a space with a measure. For instance, if measures of a C - n -tuple' components are known, the measure of the C - n -tuple itself equals to the product of these measures.

For example, if the C - n -tuple $R[XYZ] = [A \ B \ C]$ is specified in measurable attributes, and the measures of its components are equal respectively $\mu(A)$, $\mu(B)$ и $\mu(C)$, then

$$\mu(R) = \mu(A) \cdot \mu(B) \cdot \mu(C).$$

It is not that easy to calculate a measure for a C -system since intersections of some its elements can be not empty. That is why we have developed algorithms to transform any NTA structure to an orthogonal C -system that has no overlapping pairs of C - n -tuples (Kulik, 2007). Then its measure equals to the sum of the measures for the C - n -tuples comprising it.

Let us introduce two possible ways to use metric properties of NTA objects.

The first way proposes the number of elementary n -tuples as a measure based on the following regularity. Let a C - n -tuple be given in a space of attributes $P = [P_1 \ P_2 \ \dots \ P_n]$ with a finite number of their values. Then the power $|P_i|$ of every component P_i is a finite number as well, and the power (that is the number of elementary n -tuples) of the C - n -tuple P equals to $|P| = |P_1| \cdot |P_2| \cdot \dots \cdot |P_n|$. By using our orthogonalization algorithm, it is possible to count the power of every NTA object that has finite sets as its components.

In particular, this approach is applicable for a quantitative estimation of logical inference. Suppose a system of premises $A_1, A_2, \dots, A_n$ is defined in a space of attributes $X_1, X_2, \dots, X_k$ with a finite number of their values. Then the number of possible consequences from these premises equals to 2^M where $M = |U| - |A|$, $U = X_1 \times X_2 \times \dots \times X_k$ and $A = A_1 \cap_G A_2 \cap_G \dots \cap_G A_n$.

This possibility to count powers of NTA objects allows for estimating an **uncertainty measure** for a consequence B_j derived from the premises $A_1, A_2, \dots, A_n$. It looks reasonable to define such a measure as the ratio $\frac{|B_j|}{|A|}$ between the power of B_j and the power of the minimal consequence

$$A = A_1 \cap_G A_2 \cap_G \dots \cap_G A_n.$$

The second way to use metric properties of NTA objects is suitable for implementation of the logic-probabilistic analysis (LPA) for systems modeled by logical functions. LPA has been developed by I. A. Ryabinin (1976) and his colleagues. To analyze safety and reliability of such systems, LPA exploits mathematical models of the propositional calculus mostly. So, these methods are good for systems with two possible states (for example, normal functioning and failure). NTA contains methods allowing to calculate probabilities in the systems where components have more than two states (Kulik, 2007).

The presented approach corresponds to the *direct problem* of LPA, when the probability of a complex event is calculated for given probabilities of elementary events. In the *inverse problem*, it is necessary to calculate probabilities of elementary events based on the data about probabilities of certain complex events. After this, one can calculate the probabilities of other complex events. Problems solved in probabilistic logic are of this type (Kulik, 2007; Nilsson, 1986).

Let us analyze an example from (Nilsson, 1986).

Assume that a totality of the given events is defined by formulas A and $A \supset B$ of propositional calculus, besides, $p(A) = p_1$ and $p(A \supset B) = p_2$. It is necessary to estimate the probability $p(B)$ for the event B .

We will solve this problem by means of NTA. Here are two logical variables A and B , which model elementary events as well. Let us suppose probabilities of these events equal to $p(A)$ and $p(B)$ correspondingly. The statement of this problem gives $p(A) = p_1$. Now we will express the given formulas in NTA structures within the universe

$$\{0, 1\}^2: A = [\{1\} *]; B = [* \{1\}]; A \supset B = \bar{A} \vee B = \{0\} \{1\} = \begin{bmatrix} \{0\} & * \\ \{1\} & \{1\} \end{bmatrix}$$

(here the D -n-tuple modeling the formula $A \supset B$ is transformed into an orthogonal C -system).

Then the probabilistic formulas for the events A and $A \supset B$ are as follows:

$$P(A) = p_1; P(A \supset B) = (1 - p(A)) + p(A)p(B) = p_2.$$

So, we have got a system of two equations:

$$p(A) = p_1; (1 - p(A)) + p(A) p(B) = p_2$$

that allows to easily conclude

$$p(B) = \frac{p_1 + p_2 - 1}{p_1}.$$

This is the precise result for this problem while N. Nilsson (1986) gives only an interval estimate for this result, namely $p_2 + p_1 - 1 \leq p(B) \leq p_2$.

METRIZATION AND CLUSTERING OF NTA OBJECTS

For the measurable logical models discussed in the previous section, it is also possible to determine the distance by which the proximity (or remoteness) of different objects is estimated. Assume we have two NTA objects R_1 and R_2 immersed into a measurable space with a measure μ . Then the distance $d(R_1, R_2)$ between them can be calculated as:

$$d(R_1, R_2) = \mu(R_1 \Delta R_2) = \mu((R_1 \setminus R_2) \cup (R_2 \setminus R_1)). \quad (2)$$

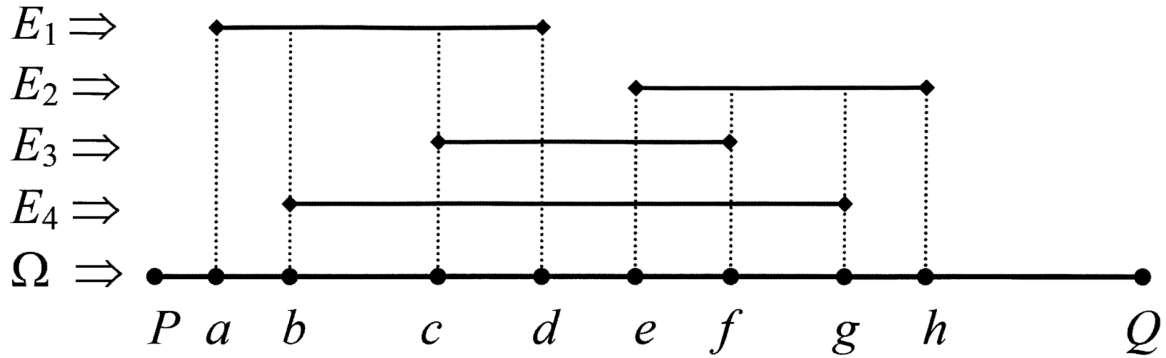
If $R_1 \subseteq R_2$, then $d(R_1, R_2) = \mu(R_2) - \mu(R_1)$. At the same time, if $R_1 \cap R_2 = \emptyset$, then $d(R_1, R_2) = \mu(R_2) + \mu(R_1)$.

The meaning of this proximity measure is that the “distance” between two NTA objects R_1 and R_2 is the smaller, the larger is the measure of the object $R_1 \cap R_2$.

Other methods to metrize NTA objects are provided for certain types of domains. These include ordered domains and interval type domains.

Consider **domains of interval type** (Kulik & Fridman, 2017). Let the domain of an attribute be defined as a closed interval Ω on a numerical axis and a finite set $E = \{E_i\}$ of closed intervals such that $\cup E_i \in \Omega$ be given. The boundaries of the intervals are given by the sets of coordinates for initial and final points. If we align all points denoting boundaries in the ascending order, we obtain a partition of the system of intervals into quanta, i.e. neighboring points and open spaces between them. The terminological difficulties associated with the fact that the set of quanta consists of dissimilar objects (points and intervals) can be circumvented by defining a point as a degenerate interval with zero measure. An example of the quantization procedure for four intervals E_1, E_2, E_3, E_4 is shown in Figure 1. For clarity, the intervals E_i are moved above the axis.

Figure 1. Quantization of a system of intervals



Here the interval Ω bounded by the points P and Q contains inner points $a, b, \dots, h$. Hence, each interval E_i contains a number of quanta. For instance, the closed interval E_3 comprises points and open intervals: $E_3 = \{c, d, e, f, (c, d), (d, e), (e, f)\}$. We can also represent E_3 as an ordered set $\{(c, d), (d, e), (e, f)\}$ of open disjoint intervals corresponding to the quanta. This way, every non-degenerate interval or non-empty result of set-theoretic operations with open intervals can be translated into an ordered set of quanta. The measure of such domains is evidently additive: the measure of an arbitrary set of quanta equals to the sum of measures for these quanta.

If an attribute domain is an **ordered set**, for example, $(a, b, c, \dots)$, we also represent it as an ordered set of quanta with names 1, 2, 3, ... If no measures of quanta are given, the measure of each quantum is assumed to be 1.

If measures of quanta are known and differ from each other (for instance, when attribute values are formed as a result of quantizing a certain set of intervals), such attributes require **normalization of intervals**, by which the average measure of all quanta becomes equal to 1. Let the measure L of an interval containing all quanta, and quantum measures l_i ($i = 1, 2, \dots, M$) be given. When normalized, **the measure of the normalized interval** is taken equal to M , and **measures of normalized quanta** (μ_i) are calculated as $\mu_i = \frac{l_i \cdot M}{L}$.

The proposed approach to defining domain measures allows to consider modeled NTA objects not only as structures immersed into a measurable space, but also as structures in a metric space. In a measurable space with domain measures, which admit using the methods outlined in the previous section, we can calculate measures of arbitrary NTA objects and their proximity by the formula (2).

If the considered model is used for solving the problem of transition to a metric space in which it is necessary to apply clustering methods, no transformations are required, it is sufficient to assume that measures of domains and their quanta defined above characterize distances.

FUTURE RESEARCH DIRECTIONS

In the authors' opinion regarding NTA development, future research ought to include the following directions: context-oriented database (knowledge base) management systems; additional means of immersing NTA structures into measure spaces and assigning NTA objects with quantitative features; modeling

dynamic intelligent systems (Vinogradov, Zhilyakova, & Osipov, 2002) within the situational approach; applications NTA methods to smart electromechanical systems (SEMS) (Kulik & Fridman, 2016).

CONCLUSION

This paper introduces a new mathematical instrument, n -tuple algebra (NTA) that implements an algebraic approach to the general theory of n -ary relations. Unlike relational algebra and theory of binary relations, NTA uses Cartesian product of sets rather than elementary n -tuples as a basic structure and is based on some generalized operations and relations. They provide logical processing of relations with different relation diagrams; these methods have no analogies in convenient theories. NTA belongs to the class of Boolean algebras. It is proved to be isomorphic to algebra of sets and supports unified implementation of various data/knowledge structures typical for artificial intelligence and many techniques of their semantic and logical analyses. Matrix properties of NTA objects allow decreasing complexity of intellectual procedures for some new structural and statistical classes of CNFs with polynomially identifiable satisfiability as well as by efficient parallelization of logical inference algorithms. NTA structures can be readily immersed into spaces with a measure. In particular, some new solutions were found for the direct and inverse problems of the logic-probabilistic analysis. Besides, the proposed approach makes it possible to use clustering methods in models of knowledge bases.

ACKNOWLEDGMENT

The authors would like to thank the Russian Foundation for Basic Researches (grants 18-29-03022, 18-07-00132, 18-01-00076, and 19-08-0079) for partial funding of this work.

REFERENCES

- Ceri, S., Gottlob, G., & Tanca, L. (1990). *Logic programming and databases*. Springer-Verlag. doi:10.1007/978-3-642-83952-8
- Codd, E. F. (1971). A Database Sublanguage Founded on the Relational Calculus. In *Proceedings of 1971 ACM-SIGFIDET Workshop on Data Description, Access and Control*. San Diego, CA: ACM.
- Kulik, B. A. (1995). A Logic Programming System Based on Cortege Algebra. *Computer and Systems Sciences International*, 33(2), 159–170.
- Kulik, B. A. (1995). New Classes of Conjunctive Normal Forms with a Polynomially Recognizable Property of Satisfiability. *Automation and Remote Control*, 56(2), 245–255.
- Kulik, B. A. (2007). N-Tuple Algebra-Based Probabilistic Logic. *Systems Analysis and Operations Research*, 46(1), 111–120.
- Kulik, B. A., & Fridman, A. Ya. (2014). Data and Knowledge Parallel Processing by Means of N-Tuple Algebra. *Journal of Engineering and Applied Sciences (Asian Research Publishing Network)*, 9(8), 1354–1357. doi:10.15622p.36.10

Kulik, B. A., & Fridman, A. Ya. (2017). *N-ary Relations for Logical Analysis of Data and Knowledge*. IGI Global.

Nilsson, N. J. (1986). Probabilistic Logic. *Artificial Intelligence*, 28(1), 71–87. doi:10.1016/0004-3702(86)90031-7

Russel, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall.

Ryabinin, I. A. (1976). *Reliability of Engineering Systems. Principles and Analysis*. Moscow: Mir.

Ullman, J. D. (1989). *Principles of Database and Knowledge-Base Systems*. Computer Science Press.

Vinogradov, A. N., Zhilyakova, L. Yu., & Osipov, G. S. (2002). Dynamic Intelligent Systems: I. Knowledge Representation and Basic Algorithms. *Computer and Systems Sciences International*, 41(6), 953–960.

Wirth, N. (1976). *Algorithms + Data Structures = Programs*. Prentice-Hall.

Zakrevskij, A. D., & Gavrilov, M. A. (1969). *Lyapas: A Programming Language for Logic and Coding Algorithms*. Academic Press.

ADDITIONAL READING

Kulik, B., Fridman, A., & Zuenko, A. (2012). Algebraic Approach to Logical Inference Implementation. *Computing and Informatics, Slovakia*, 31(6), 1295–1328.

Kulik, B. A., & Fridman, A. Ya. (2016). Logical Analysis of Data and Knowledge with Uncertainties in SEMS. In A. E. Gorodetskiy (Ed.), *Smart Electromechanical Systems* (pp. 45–59). Springer International Publishing. doi:10.1007/978-3-319-27547-5_5

Kulik, B. A., & Fridman, A. Ya. (2018). Algebra of Clusterizable Relations. In *XVIII International Conference on Data Science and Intelligent Analysis of Information*, June 4–7, 2018, Kyiv, Ukraine.

Kulik, B. A., & Fridman, A. Ya. (2018). Roles Contradictions Play in Logical Models of Metaphors and Presuppositions. In *CEUR Workshop Proceedings. Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL 2018)*, Kazan, Russia, November 1, 2018.

Kulik, B. A., & Fridman, A. Ya. (2019). Clustering Methods for Logical Models. In *Pattern Recognition and Information Processing. Proceedings of 14-th International Conference* (21–23 May 2019, Minsk, Belarus). Minsk. “Bestprint”, 2019, pp. 52–55. <http://www.ipme.ru/ipme/labs/msa/kulik/kulik.htm>

KEY TERMS AND DEFINITIONS

C-n-Tuple: An n -tuple of components defined in a certain relation diagram; domain of each component is a subset of the domain of the corresponding attribute.

C-System: A set of homotypic C - n -tuples that are denoted as a matrix in square brackets and equal to the intersection of these C - n -tuples. The rows of this matrix are C - n -tuples.

Collisions: Situations occurring during defeasible reasoning when some new knowledge (hypothesis) is inputted. Such situations can be recognized as violations of some formally expressed rules and/or limitations which control consistency and meaning content of a logical system.

D - n -Tuple: An n -tuple of components enclosed in reversed square brackets and equal to a diagonal C -system with the same diagonal components.

D -System: A set of homotypic D - n -tuples equal to the intersection of these D - n -tuples.

Generalized Operations and Relations: They differ from similar operations and relations of algebra of sets by the only feature: NTA objects (operands) are reduced to the same relation diagram before executing these operations or checking the relations.

N -Tuple Algebra: An algebraic system whose support is an arbitrary set of n -ary relations expressed by specific structures, namely, C - n -tuple, C -system, D - n -tuple, and D -system. These structures provide a compact expression for sets of elementary n -tuples.

Intelligent Constructing Exact Tolerance Limits for Prediction of Future Outcomes Under Parametric Uncertainty

Nicholas Nechval

University of Latvia, Latvia

INTRODUCTION

The logical purpose for a statistical tolerance limit (where the coverage value γ is the percentage of the future process outcomes to be captured by the prediction, and the confidence level $(1-\alpha)$ is the proportion of the time we hope to capture that percentage γ) is to predict future outcomes for some production process which is treated as process, say, with stochastic variation of a product lifetime. The applications of tolerance limits (intervals) are varied. They included clinical and industrial applications, including quality control, applications to environmental monitoring, to the assessment of agreement between two methods or devices, and applications in industrial hygiene. For example, such tolerance limits are required, when planning life tests, engineers may need to predict the number of failures that will occur by the end of the test or to predict the amount of time that it will take for a specified number of units to fail. Tolerance limits of the type mentioned above are considered in this article, which presents a new technique for constructing exact statistical (lower and upper) tolerance limits on outcomes (for example, on order statistics) in future samples. Attention is restricted to the extreme-value and two-parameter Weibull distributions under parametric uncertainty (when both parameters are unknown). The technique used here emphasizes pivotal quantities relevant for obtaining tolerance factors and is applicable whenever the statistical problem is invariant under a group of transformations that acts transitively on the parameter space. It does not require the construction of any tables and is applicable whether the experimental data are complete or Type II censored. The exact tolerance limits on order statistics associated with sampling from underlying distributions can be found easily and quickly making tables, simulation, Monte Carlo estimated percentiles, special computer programs, and approximation unnecessary. The proposed technique is based on a probability transformation and pivotal quantity averaging. It does not in need to make any assumption concerning the statistical functional form for the tolerance limit, is conceptually simple and easy to use. The scientific literature does not contain an analytical methodology for constructing exact γ -content tolerance limits with expected $(1-\alpha)$ -confidence on future order statistics coming from an extreme-value or Weibull distribution. One reason is that the theoretical concept and computational complexity of the tolerance limits is significantly more difficult than that of the standard confidence and prediction limits. However, in the literature there are several known methods for constructing $(1-\alpha)$ -prediction limits (in terms of this article, tolerance limits with expected $(1-\alpha)$ -confidence) on future order statistics coming from the two-parameter Weibull distribution. Therefore, finally, we give numerical examples, where the $(1-\alpha)$ -prediction limits obtained by using the known methods are compared with the results obtained through the proposed analytical methodology, which is illustrated in terms of the extreme-value and two-parameter Weibull distributions. Analytical formulas for the tolerance limits are available in the scientific literature for only simple cases, for example, for the upper or lower tolerance

DOI: 10.4018/978-1-7998-3479-3.ch049

limit for a univariate normal population. Thus it becomes necessary to use new methods in order to derive exact statistical tolerance limits for many populations. The proposed, in this article, technique of intelligent constructing exact statistical γ -content tolerance limits with expected $(1-\alpha)$ -confidence, which are obtained here in terms of the two-parameter Weibull and extreme-value distributions, represents a novelty in the theory of statistical decisions.

BACKGROUND

In this paper, two types of statistical tolerance limits are defined: i) γ -content tolerance limits with expected $(1-\alpha)$ -confidence on future outcomes, ii) tolerance limits with expected $(1-\alpha)$ -confidence on future outcomes. To be specific, let γ denote a proportion between 0 and 1. Then one-sided statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence is determined to capture a proportion γ or more of the population, with a given expected confidence level $1-\alpha$. For example, an upper statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence on future outcomes from a univariate population is such that with the given expected confidence level $1-\alpha$, a specified proportion γ or more of the population will fall below the limit. A lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence satisfies similar conditions. An upper statistical tolerance limit with expected $(1-\alpha)$ -confidence is determined so that the expected proportion of the population failing below the limit is $(1-\alpha)$. A lower statistical tolerance limit with expected $(1-\alpha)$ -confidence satisfies similar conditions. The statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence seems to be more useful than the statistical tolerance limit with expected $(1-\alpha)$ -confidence but is relatively difficult to construct.

The logical purpose for a tolerance limit must be the prediction of future outcomes for some (say, stochastic) process. Tolerance (prediction) limits enjoy a fairly rich history in the scientific literature and have a very important role in engineering and manufacturing applications. Patel (1986) provides a review (which was fairly comprehensive at the time of publication) of tolerance intervals (limits) for many distributions as well as a discussion of their relation with confidence intervals (limits) for percentiles. Dunsmore (1978) and Guenther, Patil, and Uppuluri (1976) both discuss 2-parameter exponential tolerance intervals (limits) and the estimation procedure in greater detail. Engelhardt and Bain (1978) discuss how to modify the formulas when dealing with type II censored data. Guenther (1972) and Hahn and Meeker (1991) discuss how one-sided tolerance limits can be used to obtain approximate two-sided tolerance intervals by applying Bonferroni's inequality. In Nechval et al. (1999, 2004, 2014a, 2016a, 2016b, 2016c, 2017, 2018a, 2019), the exact statistical tolerance and prediction limits are discussed under parametric uncertainty of underlying models.

In contrast to other statistical limits commonly used for statistical inference, the γ -content tolerance limits with expected $(1-\alpha)$ -confidence (especially for the order statistics) are used relatively rarely. One reason is that the theoretical concept and computational complexity of the γ -content tolerance limits with expected $(1-\alpha)$ -confidence is significantly more difficult than that of the standard confidence and prediction limits. Thus it becomes necessary to use the innovative approaches which will allow one to construct tolerance limits on future order statistics for many populations.

FOCUS OF THE ARTICLE

Problem Statement

The problem can be stated more formally as follows. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable X from a sample of size n from a distribution with a probability density function $f_{\theta}(x)$ (distribution function $F_{\theta}(x)$, survival function $\bar{F}_{\theta}(x) = 1 - F_{\theta}(x)$) and S be any statistic (say, sufficient statistic or maximum likelihood estimator) obtained from the experimental random sample $X_1 \leq \dots \leq X_r$ and let a random variable Y (in a future random sample $Y_1, \dots, Y_m$) has the same distribution with the probability density function $f_{\theta}(y)$ (distribution function $F_{\theta}(y)$, survival function $\bar{F}_{\theta}(y) = 1 - F_{\theta}(y)$), where a parameter θ (in general, vector) is common to both distributions and it is assumed that some or all numerical values of components of the parametric vector θ are unspecified.

On the basis of the experimental random sample $X_1 \leq \dots \leq X_r$ we wish to make a prediction about a future outcome of Y_k (k th order statistic, $1 \leq k \leq m$, in a future random sample of m ordered observations $Y_1 \leq \dots \leq Y_m$), usually in the form of one-sided statistical tolerance limits on future outcomes of Y_k (lower γ -content tolerance limit L_k with expected $(1-\alpha)$ -confidence and upper γ -content tolerance limit U_k with expected $(1-\alpha)$ -confidence). That is, if L_k and U_k are functions of S , then $L_k \equiv L_k(S)$ is a lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence on future outcomes of the k th order statistic Y_k if

$$E_{\theta} \left\{ \Pr \left(\int_{L_k(S)}^{\infty} g_{\theta}(y_k) dy_k \geq \gamma \right) \right\} = E_{\theta} \left\{ \Pr \left(\bar{G}_{\theta}(L_k(S)) \geq \gamma \right) \right\} = 1 - \alpha, \quad (1)$$

and $U_k \equiv U_k(S)$ is an upper statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence on future outcomes of the k th order statistic Y_k if

$$E_{\theta} \left\{ \Pr \left(\int_0^{U_k(S)} g_{\theta}(y_k) dy_k \geq \gamma \right) \right\} = E_{\theta} \left\{ \Pr \left(G_{\theta}(U_k(S)) \geq \gamma \right) \right\} = 1 - \alpha, \quad (2)$$

where

$$g_{\theta}(y_k) = \frac{1}{B(k, m-k+1)} [F_{\theta}(y_k)]^{k-1} [1 - F_{\theta}(y_k)]^{m-k} f_{\theta}(y_k) \quad (3)$$

is the probability density function of the k th order statistic Y_k ,

$$\begin{aligned} G_{\theta}(y_k) &= \Pr(Y_k \leq y_k) = \sum_{i=k}^m \binom{m}{i} [F_{\theta}(y_k)]^i [1 - F_{\theta}(y_k)]^{m-i} = \sum_{i=k}^m \binom{m}{i} [1 - \bar{F}_{\theta}(y_k)]^i [\bar{F}_{\theta}(y_k)]^{m-i} \\ &= \sum_{i=k}^m \binom{m}{i} \sum_{j=0}^i \binom{i}{j} (-1)^j [\bar{F}_{\theta}(y_k)]^{m-i+j} = \int_0^{F_{\theta}(y_k)} \frac{1}{B(k, m-k+1)} \tau^{k-1} (1-\tau)^{m-k} d\tau \end{aligned} \quad (4)$$

is the probability distribution function of the k th order statistic Y_k . It can be shown that

$$\frac{dG_{\theta}(y_k)}{dy_k} = g_{\theta}(y_k). \quad (5)$$

Further, $L_k^{\circ} \equiv L_k^{\circ}(S)$ is a lower statistical tolerance limit with expected $(1-\alpha)$ -confidence on future outcomes of the k th order statistic Y_k from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ if it satisfies

$$E_{\theta} \left\{ \Pr(Y_k > L_k^{\circ}(S)) \right\} = E_{\theta} \left\{ \int_{L_k^{\circ}(S)}^{\infty} g_{\theta}(y_k) dy_k \right\} = E_{\theta} \left\{ \bar{G}_{\theta}(L_k^{\circ}(S)) \right\} = 1 - \alpha. \quad (6)$$

$U_k^{\circ} \equiv U_k^{\circ}(S)$ is an upper statistical tolerance limit with expected $(1-\alpha)$ -confidence on future outcomes of the k th order statistic Y_k from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ if it satisfies

$$E_{\theta} \left\{ \Pr(Y_k \leq U_k^{\circ}(S)) \right\} = E_{\theta} \left\{ \int_0^{U_k^{\circ}(S)} g_{\theta}(y_k) dy_k \right\} = E_{\theta} \left\{ G_{\theta}(U_k^{\circ}(S)) \right\} = 1 - \alpha. \quad (7)$$

In this article, a new technique for intelligent constructing the statistical γ -content tolerance limits with expected $(1-\alpha)$ -confidence as well as the statistical tolerance limits with expected $(1-\alpha)$ -confidence on order statistics in future samples is proposed. For illustration, the extreme-value and Weibull distributions are considered.

Extreme-Value Distribution

This distribution is used in many research fields including, among others, life testing and water resource management. This is the so-called first asymptotic distribution of extreme values, hereafter referred to simply as the extreme-value distribution. The distribution is extensively used in a number of areas as a lifetime distribution and sometimes referred to as the Gumbel distribution, after E. J. Gumbel, who had pioneered its use (Gumbel, 1958).

Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable X from a sample of size n from an extreme-value distribution with the pdf (probability density function),

$$f_{\theta}(x) = \frac{1}{\theta_2} \exp\left(\frac{x - \theta_1}{\theta_2}\right) \exp\left[-\exp\left(\frac{x - \theta_1}{\theta_2}\right)\right], \quad -\infty < x < \infty, \quad (8)$$

and cdf (cumulative distribution function),

$$F_{\theta}(x) = 1 - \exp\left[-\exp\left(\frac{x - \theta_1}{\theta_2}\right)\right], \quad -\infty < x < \infty, \quad (9)$$

indexed by location and scale parameters θ_1 and θ_2 , where $\theta = (\theta_1, \theta_2)$. It is assumed that the parameters $\theta_1 (-\infty < \theta_1 < \infty)$ and $\theta_2 > 0$ are unknown.

In Type II censoring, which is of primary interest here, the number of survivors are fixed and X_r is a random variable. The MLE's $\hat{\theta}_1$ and $\hat{\theta}_2$ of the parameters θ_1 and θ_2 , respectively, are solutions of

$$\hat{\theta}_1 = \ln \left[r^{-1} \left(\sum_{i=1}^r \exp \left(\frac{x_i}{\hat{\theta}_2} \right) + (n-r) \exp \left(\frac{x_r}{\hat{\theta}_2} \right) \right) \right]^{\hat{\theta}_2}, \quad (10)$$

$$\hat{\theta}_2 = \left[\sum_{i=1}^r x_i \exp \left(\frac{x_i}{\hat{\theta}_2} \right) + (n-r) x_r \exp \left(\frac{x_r}{\hat{\theta}_2} \right) \right] \left[\sum_{i=1}^r \exp \left(\frac{x_i}{\hat{\theta}_2} \right) + (n-r) \exp \left(\frac{x_r}{\hat{\theta}_2} \right) \right]^{-1} - \frac{1}{r} \sum_{i=1}^r x_i. \quad (11)$$

In terms of the extreme-value distribution variates, we have that

$$W_1 = \frac{\hat{\theta}_1 - \theta_1}{\theta_2}, \quad W_2 = \frac{\hat{\theta}_2}{\theta_2}, \quad W_3 = \frac{\hat{\theta}_1 - \theta_1}{\hat{\theta}_2} \quad (12)$$

are pivotal quantities. The probability density functions of the pivotal quantities do not depend on the parameters. It can be shown that the joint pdf of the pivotal quantities

$$W_1 = \frac{\hat{\theta}_1 - \theta_1}{\theta_2}, \quad W_2 = \frac{\hat{\theta}_2}{\theta_2}, \quad (13)$$

conditional on fixed

$$\mathbf{S}^{(r)} = (S_r, \dots, S_r), \quad (14)$$

where

$$S_i = \frac{X_i - \hat{\theta}_1}{\hat{\theta}_2}, \quad i = 1, \dots, r, \quad (15)$$

are ancillary statistics, any $r-2$ of which form a functionally independent set, $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimates for θ_1 and θ_2 , respectively, based on the first r ordered observations ($X_1 \leq \dots \leq X_r$) from a sample of size n from the extreme-value distribution (8), which can be found from solution of (10) and (11), is given by

$$\begin{aligned}
 & f_n(w_1, w_2 | \mathbf{s}^{(r)}) \\
 &= \left(\frac{(e^{w_1})^r}{\Gamma(r)} \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^r \exp \left(-e^{w_1} \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right] \right) \right) \\
 &\times \left(\frac{1}{v(\mathbf{s}^{(r)})} w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r} \right) \\
 &= f_n(w_1 | \mathbf{s}^{(r)}, w_2) f_n(w_2 | \mathbf{s}^{(r)}), w_1 \in (-\infty, \infty), w_2 \in (0, \infty),
 \end{aligned} \tag{16}$$

where

$$\begin{aligned}
 f_n(w_1 | \mathbf{s}^{(r)}, w_2) &= \frac{(e^{w_1})^r}{\Gamma(r)} \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^r \\
 &\exp \left(-e^{w_1} \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right] \right), w_1 \in (-\infty, \infty).
 \end{aligned} \tag{17}$$

$$f_n(w_2 | \mathbf{s}^{(r)}) = \frac{1}{v(\mathbf{s}^{(r)})} w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r}, w_2 \in (0, \infty), \tag{18}$$

$$v(\mathbf{s}^{(r)}) = \int_0^\infty w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r} dw_2 \tag{19}$$

is the normalizing constant. If a pivotal quantity is given by

$$W = e^{W_1} \left[\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2) \right], \tag{20}$$

it follows from (17) that

$$W \sim g_r(w) = \frac{1}{\Gamma(r)} w^{r-1} \exp(-w), w \in (0, \infty). \tag{21}$$

Two-Parameter Weibull Distribution

The two-parameter Weibull distribution is one of the most widely used life distributions in reliability analysis. This distribution is very flexible, and can, through an appropriate choice of parameters, model many types of failure rate behaviors. It has wide applications in diverse disciplines.

Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable X from a sample of size n from a two-parameter Weibull distribution with the pdf,

$$f_{\theta}(\underline{x}) = \frac{\delta}{\beta} \left(\frac{\underline{x}}{\beta} \right)^{\delta-1} \exp \left[- \left(\frac{\underline{x}}{\beta} \right)^{\delta} \right], \quad \underline{x} > 0, \quad \beta > 0, \quad \delta > 0, \quad (22)$$

and cdf,

$$F_{\theta}(\underline{x}) = 1 - \exp \left[- \left(\frac{\underline{x}}{\beta} \right)^{\delta} \right], \quad \underline{x} > 0, \quad \beta > 0, \quad \delta > 0, \quad (23)$$

indexed by scale and shape parameters β and δ , where $\theta = (\beta, \delta)$. It is assumed that the parameters β and δ are unknown. This distribution is directly related to the extreme-value distribution by the easily shown fact that if X has a Weibull distribution (22), then $X = \ln X$ has an extreme-value distribution with $\theta_1 = \ln \beta$ and $\theta_2 = \delta^{-1}$. In analyzing data it is often convenient to work with log times, the extreme-value distribution arises when lifetimes are taken to be Weibull distributed. The MLE's of the Weibull parameters β and δ are $\hat{\beta} = \exp \hat{\theta}_1$ and $\hat{\delta} = \hat{\theta}_2^{-1}$. If desired, the maximum likelihood equations (10) and (11) can be written in Weibull form and solved directly from the start. The equations are

$$\hat{\beta} = \left(r^{-1} \left(\sum_{i=1}^r x_i^{\hat{\delta}} + (n-r) \underline{x}_r^{\hat{\delta}} \right) \right)^{1/\hat{\delta}}, \quad (24)$$

$$\hat{\delta} = \left[\left(\sum_{i=1}^r \underline{x}_i^{\hat{\delta}} \ln \underline{x}_i + (n-r) \underline{x}_r^{\hat{\delta}} \ln \underline{x}_r \right) \left(\sum_{i=1}^r \underline{x}_i^{\hat{\delta}} + (n-r) \underline{x}_r^{\hat{\delta}} \right)^{-1} - \frac{1}{r} \sum_{i=1}^r \ln \underline{x}_i \right]^{-1}. \quad (25)$$

In terms of the Weibull variates, we have that

$$V_1 = \left(\frac{\hat{\beta}}{\beta} \right)^{\delta}, \quad V_2 = \frac{\delta}{\hat{\delta}}, \quad V_3 = \left(\frac{\hat{\beta}}{\beta} \right)^{\hat{\delta}} \quad (26)$$

are pivotal quantities. The probability density functions of the pivotal quantities do not depend on the parameters. It can be shown that the joint pdf of the pivotal quantities

$$V_1 = \left(\frac{\hat{\beta}}{\beta} \right)^\delta, \quad V_2 = \frac{\delta}{\hat{\delta}}, \quad (27)$$

conditional on fixed

$$\mathbf{Z}^{(r)} = (Z_1, \dots, Z_r), \quad (28)$$

where

$$Z_i = \left(\frac{X_i}{\beta} \right)^\delta, \quad i = 1, \dots, r, \quad (29)$$

are ancillary statistics, any $r-2$ of which form a functionally independent set, $\hat{\beta}$ and $\hat{\delta}$ are the maximum likelihood estimates for β and δ , respectively, based on the first r ordered observations $X_1 \leq \dots \leq X_r$ from a sample of size n from the two-parameter Weibull distribution (22), which can be found from solution of (24) and (25), is given by

$$\begin{aligned} f_n(v_1, v_2 | \mathbf{z}^{(r)}) &= \left(\frac{v_1^{r-1}}{\Gamma(r)} \left[\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right]^r \exp \left(-v_1 \left[\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right] \right) \right) \\ &\times \left(\frac{1}{g(\mathbf{z}^{(r)})} v_2^{r-2} \prod_{i=1}^r z_i^{v_2} \left[\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right]^{-r} \right) = f_n(v_1 | \mathbf{z}^{(r)}, v_2) f_n(v_2 | \mathbf{z}^{(r)}), \end{aligned} \quad (30)$$

$v_1 \in (0, \infty), \quad v_2 \in (0, \infty),$

where

$$f_n(v_1 | \mathbf{z}^{(r)}, v_2) = \frac{1}{\Gamma(r)} \left[\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right]^r v_1^{r-1} \exp \left(-v_1 \left[\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right] \right), \quad v_1 \in (0, \infty), \quad (31)$$

$$f_n(v_2 | \mathbf{z}^{(r)}) = \frac{1}{g(\mathbf{z}^{(r)})} v_2^{r-2} \prod_{i=1}^r z_i^{v_2} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)^{-r}, \quad v_2 \in (0, \infty), \quad (32)$$

$$g(\mathbf{z}^{(r)}) = \int_0^\infty v_2^{r-2} \prod_{i=1}^r z_i^{v_2} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)^{-r} dv_2 \quad (33)$$

is the normalizing constant. If a pivotal quantity is given by

$$V = V_1 \left[\sum_{i=1}^r Z_i^{V_2} + (n-r)Z_r^{V_2} \right], \quad (34)$$

it follows from (31) that

$$V \sim g_r(v) = \frac{1}{\Gamma(r)} v^{r-1} \exp(-v), \quad v \in (0, \infty). \quad (35)$$

CONSTRUCTING STATISTICAL γ -CONTENT TOLERANCE LIMITS WITH EXPECTED $(1-\alpha)$ -CONFIDENCE

Constructing Lower Statistical γ -Content Tolerance Limit with Expected $(1-\alpha)$ -Confidence

Theorem 1. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\ln \underline{X})$ from a sample of size n from an extreme-value distribution defined by the probability density function (8). Then a lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $L_k \equiv L_k(S)$, on future outcomes of the k th order statistic $Y_k(=\ln Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (8), which satisfies (1) is given by

$$L_k = \hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{L_k} = \ln \underline{L}_k, \quad (36)$$

where η_{L_k} is a tolerance factor determined by

$$\eta_{L_k} = \arg \left(\int_0^\infty \int_0^\infty \frac{\ln(1-q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right)}{[\eta_{L_k}]^{w_2}} \frac{1}{\Gamma(r)} w^{r-1} \exp(-w) \right. \\ \left. \times \frac{1}{v(s^{(r)})} w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r} dw dw_2 = 1 - \alpha \right); \quad (37)$$

the maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of the parameters θ_1 and θ_2 are determined from (10) and (11), respectively; the ancillary statistics S_i , $i=1, \dots, r$, are given by (15); $q_{1-\gamma}$ is a quantile of the beta distribution satisfying

$$\int_0^{q_{1-\gamma}} \frac{1}{B(k, m-k+1)} \tau^{k-1} (1-\tau)^{m-k} d\tau = 1 - \gamma. \quad (38)$$

Proof. It follows from (3) and (4) that $\Pr\left(\int_{L_k}^{\infty} g_{\theta}(y_k) dy_k \geq \gamma\right)$

$$\begin{aligned}
 &= \Pr\left(\bar{G}_{\theta}(L_k) \geq \gamma\right) = \Pr\left(1 - G_{\theta}(L_k) \geq \gamma\right) = \Pr\left(1 - \int_0^{F_{\theta}(L_k)} \frac{1}{B(k, m-k+1)} \tau^{k-1} (1-\tau)^{m-k} d\tau \geq \gamma\right) \\
 &= \Pr\left(\int_0^{F_{\theta}(L_k)} \frac{1}{B(k, m-k+1)} \tau^{k-1} (1-\tau)^{m-k} d\tau \leq 1-\gamma\right) = \Pr\left(F_{\theta}(L_k) \leq q_{1-\gamma}\right) \\
 &= \Pr\left(1 - \exp\left[-\exp\left(\frac{L_k - \theta_1}{\theta_2}\right)\right] \leq q_{1-\gamma}\right) = \Pr\left(\exp\left[-\exp\left(\frac{L_k - \theta_1}{\theta_2}\right)\right] \geq 1 - q_{1-\gamma}\right) = \Pr\left(e^{W_1} \leq \frac{\ln(1 - q_{1-\gamma})^{-1}}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2)\right]^{W_2}}\right) \\
 &= \Pr\left(e^{W_1} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2)\right) \leq \frac{\ln(1 - q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2)\right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2)\right]^{W_2}}\right) \\
 &= \Pr\left(W \leq \frac{\ln(1 - q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2)\right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2)\right]^{W_2}}\right). \tag{39}
 \end{aligned}$$

Using pivotal quantity averaging, it follows from (1) and (39) that

$$\begin{aligned}
 E_{\theta} \left\{ \Pr\left(\int_{L_k(S)}^{\infty} g_{\theta}(y_k) dy_k \geq \gamma\right) \right\} &= E \left\{ \Pr\left(W \leq \frac{\ln(1 - q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2)\right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2)\right]^{W_2}}\right) \right\} \\
 &= E \left\{ \int_0^{\frac{\ln(1 - q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2)\right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2)\right]^{w_2}}} g_r(w) dw \right\} = \int_0^{\infty} \int_0^{\frac{\ln(1 - q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2)\right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2)\right]^{w_2}}} g_r(w) f_n(v_2 | \mathbf{z}^{(r)}) dw dv_2
 \end{aligned}$$

$$\begin{aligned}
&= \int_0^\infty \int_0^\infty \frac{\ln(1-q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2) \right]^{w_2}} \frac{1}{\Gamma(r)} w^{r-1} \exp(-w) \\
&\times \frac{1}{v(\mathbf{s}^{(r)})} w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r} dw dw_2 = 1 - \alpha. \quad (40)
\end{aligned}$$

It follows from (40) that

$$\begin{aligned}
L_k = \arg \left(\int_0^\infty \int_0^\infty \frac{\ln(1-q_{1-\gamma})^{-1} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right)}{\left[\exp((L_k - \hat{\theta}_1)/\hat{\theta}_2) \right]^{w_2}} \frac{1}{\Gamma(r)} w^{r-1} \exp(-w) \right. \\
\left. \times \frac{1}{v(\mathbf{s}^{(r)})} w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r} dw dw_2 = 1 - \alpha \right). \quad (41)
\end{aligned}$$

Assuming that

$$\exp\left(\frac{L_k - \hat{\theta}_1}{\hat{\theta}_2}\right) = \eta_{L_k}, \quad (42)$$

we have (36). This completes the proof.

Theorem 2. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\exp X)$ from a sample of size n from a two-parameter Weibull distribution defined by the probability density function (22). Then a lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k \equiv L_k(\mathcal{S})$, on future outcomes of the k th order statistic $Y_k(\exp Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (22), which satisfies (1) is given by

$$\underline{L}_k = \eta_{\underline{L}_k}^{1/\hat{\delta}} \hat{\beta} = \exp L_k, \quad (43)$$

where $\eta_{\underline{L}_k}$ is a tolerance factor determined by

$$\eta_{L_k} = \arg \left(\int_0^\infty \int_0^\infty \frac{\ln(1-q_{1-\gamma})^{-1} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)}{[\eta_{L_k}]^{v_2}} \frac{1}{\Gamma(r)} v^{r-1} \exp(-v) \frac{1}{g(\mathbf{z}^{(r)})} v_2^{r-2} \prod_{i=1}^r z_i^{v_2} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)^{-r} dv dv_2 = 1 - \alpha \right); \quad (44)$$

the maximum likelihood estimates $\hat{\beta}$ and $\hat{\delta}$ of the parameters β and δ are determined from (24) and (25), respectively; the ancillary statistics Z_i , $i=1, \dots, r$, are given by (29); $q_{1-\gamma}$ is a quantile of the beta distribution satisfying (38).

Proof. The proof is similar to that of Theorem 1 and so it is omitted here.

Inference 1 (for the tolerance factors η_{L_k} and η_{U_k}). It follows from (42) and (43) that

$$\eta_{L_k} = \left(\frac{L_k}{\hat{\beta}} \right)^{\hat{\delta}} = \left(\frac{\exp L_k}{\exp \hat{\theta}_1} \right)^{1/\hat{\theta}_2} = \exp \left(\frac{L_k - \hat{\theta}_1}{\hat{\theta}_2} \right) = \eta_{L_k}. \quad (45)$$

Constructing Upper Statistical γ -Content Tolerance Limit with Expected $(1-\alpha)$ -Confidence

Theorem 3. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\ln X)$ from a sample of size n from an extreme-value distribution defined by the probability density function (8). Then an upper statistical γ -content tolerance limit with $(1-\alpha)$ -confidence, $U_k \equiv U_k(S)$, on future outcomes of the k th order statistic $Y_k(=\ln Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (8), which satisfies (2) is given by

$$U_k = \hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{U_k} = \ln U_k, \quad (46)$$

where η_{U_k} is a tolerance factor determined by

$$\eta_{U_k} = \arg \left(\int_0^\infty \int_0^\infty \frac{\ln(1-q_\gamma)^{-1} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right)}{[\eta_{U_k}]^{w_2}} \frac{1}{\Gamma(r)} w^{r-1} \exp(-w) \right. \\ \left. \times \frac{1}{v(\mathbf{s}^{(r)})} w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2) \left[\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right]^{-r} dw dw_2 = \alpha \right); \quad (47)$$

the maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of the parameters θ_1 and θ_2 are determined from (10) and (11), respectively; the ancillary statistics S_i , $i=1, \dots, r$, are given by (15); q_γ is a quantile of the beta distribution satisfying (38).

Proof. The upper statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $U_k \equiv U_k(S)$, is obtained from a lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $L_k \equiv L_k(S)$, by replacing γ by $1-\gamma$, and $1-\alpha$ by α . This completes the proof.

Theorem 4. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\exp X)$ from a sample of size n from a two-parameter Weibull distribution defined by the probability density function (22). Then an upper statistical γ -content tolerance limit with $(1-\alpha)$ -confidence, $\underline{U}_k \equiv \underline{U}_k(S)$, on future outcomes of the k th order statistic $Y_k(=\exp Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (22), which satisfies (2) is given by

$$\underline{U}_k = \eta_{\underline{U}_k}^{1/\hat{\delta}} \hat{\beta} = \exp U_k, \quad (48)$$

where $\eta_{\underline{U}_k}$ is a tolerance factor determined by

$$\eta_{\underline{U}_k} = \arg \left[\int_0^\infty \int_0^\infty \frac{\ln(1-q_\gamma)^{-1} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)}{[\eta_{\underline{U}_k}]^{v_2}} \frac{1}{\Gamma(r)} v^{r-1} \exp(-v) \frac{1}{g(\mathbf{z}^{(r)})} v_2^{r-2} \prod_{i=1}^r z_i^{v_2} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)^{-r} dv dv_2 = \alpha \right]. \quad (49)$$

Proof. The upper statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{U}_k \equiv \underline{U}_k(S)$, is obtained from a lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $L_k \equiv L_k(S)$, by replacing γ by $1-\gamma$, and $1-\alpha$ by α . This completes the proof.

Inference 2 (for the tolerance factors $\eta_{\underline{U}_k}$ and η_{U_k}). It follows from (46) and (48) that

$$\eta_{\underline{U}_k} = \left(\frac{\underline{U}_k}{\hat{\beta}} \right)^{\hat{\delta}} = \left(\frac{\exp U_k}{\exp \hat{\theta}_1} \right)^{1/\hat{\theta}_2} = \exp \left(\frac{U_k - \hat{\theta}_1}{\hat{\theta}_2} \right) = \eta_{U_k}. \quad (50)$$

CONSTRUCTING STATISTICAL TOLERANCE LIMITS WITH EXPECTED (1- α)-CONFIDENCE

Constructing Lower Statistical Tolerance Limit with Expected (1- α)-Confidence

Theorem 5. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\ln X)$ from a sample of size n from an extreme-value distribution defined by the probability density function (8). Then a lower statistical tolerance limit with expected (1- α)-confidence, $L_k^\circ \equiv L_k^\circ(S)$, on future outcomes of the k th order statistic $Y_k(=\ln L_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (8), which satisfies (6) is given by

$$L_k^\circ = \hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{L_k^\circ} = \ln \underline{L}_k^\circ, \quad (51)$$

where $\eta_{L_k^\circ}$ is a tolerance factor determined by

$$\eta_{L_k^\circ} = \arg \left(\sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{v(\mathbf{s}^{(r)})} \int_0^\infty \frac{w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2)}{\left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) + [\eta_{L_k^\circ}]^{w_2} \right)^r (m-l+j)} dw_2 = 1 - \alpha \right). \quad (52)$$

Proof. It follows from (3) and (4) that

$$\begin{aligned} \Pr(Y_k > L_k^\circ) &= \int_{L_k^\circ}^\infty g_\theta(y_k) dy_k = \bar{G}_\theta(L_k^\circ) = \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j [\bar{F}_\theta(L_k^\circ)]^{m-l+j} \\ &= \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \left(\exp \left[-\exp \left(\frac{L_k^\circ - \theta_1}{\theta_2} \right) \right] \right)^{m-l+j} = \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \\ &\times \left(\exp \left[-e^{w_1} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2) \right) \right] \left(\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) \right)^{W_2} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2) \right)^{-1} \right)^{m-l+j} \\ &= \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \left(\exp \left[-W \left(\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) \right)^{W_2} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2) \right)^{-1} (m-l+j) \right] \right). \end{aligned} \quad (53)$$

Using pivotal quantity averaging, it follows from (6) and (53) that $E_\theta \left\{ \Pr(Y_k > L_k^\circ(S)) \right\}$

$$\begin{aligned}
&= E \left\{ \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \left[\exp \left[-W \left(\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) \right)^{w_2} \left(\sum_{i=1}^r \exp(S_i W_2) + (n-r) \exp(S_r W_2) \right)^{-1} (m-l+j) \right] \right] \right\} \\
&= \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \\
&\quad \times \int_0^\infty \int_0^\infty \exp \left[-w \left(\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) \right)^{w_2} \left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) \right)^{-1} (m-l+j) \right] g_r(w) f_n(w_2 | \mathbf{s}^{(r)}) dw dw_2 \\
&= \sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{v(\mathbf{s}^{(r)})} \int_0^\infty \frac{w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2)}{\left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) + \left(\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) \right)^{w_2} (m-l+j) \right)^r} dw_2 = 1 - \alpha.
\end{aligned} \tag{54}$$

It follows from (54) that

$$\begin{aligned}
L_k^\circ = \arg \left(\sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{v(\mathbf{s}^{(r)})} \right. \\
\left. \times \int_0^\infty \frac{w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2)}{\left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) + \left(\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) \right)^{w_2} (m-l+j) \right)^r} dw_2 = 1 - \alpha \right).
\end{aligned} \tag{55}$$

Assuming that

$$\exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) = \eta_{L_k^\circ}, \tag{56}$$

we have (51). This completes the proof.

Theorem 6. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(= \exp X)$ from a sample of size n from a two-parameter Weibull distribution defined by the probability density function

(22). Then a lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k^\circ \equiv \underline{L}_k^\circ(S)$, on future outcomes of the k th order statistic $Y_k(=\exp Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (22), which satisfies (6) is given by

$$\underline{L}_k^\circ = \eta_{\underline{L}_k^\circ}^{1/\hat{\delta}} \hat{\beta} = \exp L_k^\circ, \quad (57)$$

where $\eta_{\underline{L}_k^\circ}$ is a tolerance factor determined by

$$\eta_{\underline{L}_k^\circ} = \arg \left(\sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{g(\mathbf{z}^{(r)})} \int_0^\infty \frac{v_2^{r-2} \prod_{i=1}^r z_i^{v_2}}{\left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} + [\eta_{\underline{L}_k^\circ}]^{v_2} (m-l+j) \right)^r} dv_2 = 1-\alpha \right). \quad (58)$$

Proof. The proof is similar to that of Theorem 5 and so it is omitted here.

Inference 3 (for the tolerance factors $\eta_{\underline{L}_k^\circ}$ and $\eta_{\underline{L}_k^\circ}$). It follows from (56) and (57) that

$$\eta_{\underline{L}_k^\circ} = \left(\frac{\underline{L}_k^\circ}{\hat{\beta}} \right)^{\hat{\delta}} = \left(\frac{\exp L_k^\circ}{\exp \hat{\theta}_1} \right)^{1/\hat{\theta}_2} = \exp \left(\frac{L_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) = \eta_{\underline{L}_k^\circ}. \quad (59)$$

Constructing Upper Statistical Tolerance Limit With Expected $(1-\alpha)$ -Confidence

Theorem 7. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\ln X)$ from a sample of size n from an extreme-value distribution defined by the probability density function (8). Then an upper statistical tolerance limit with expected $(1-\alpha)$ -confidence, $U_k^\circ \equiv U_k^\circ(S)$, on future outcomes of the k th order statistic $Y_k(=\ln Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (8), which satisfies (7) is given by

$$U_k^\circ = \hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{U_k^\circ} = \ln \underline{U}_k^\circ, \quad (60)$$

where $\eta_{U_k^\circ}$ is a tolerance factor determined by

$$\eta_{U_k^\circ} = \arg \left(\sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{v(\mathbf{s}^{(r)})} \int_0^\infty \frac{w_2^{r-2} \prod_{i=1}^r \exp(s_i w_2)}{\left(\sum_{i=1}^r \exp(s_i w_2) + (n-r) \exp(s_r w_2) + [\eta_{U_k^\circ}]^{w_2} (m-l+j) \right)^r} dw_2 = \alpha \right). \quad (61)$$

Proof. The upper statistical tolerance limit with expected $(1-\alpha)$ -confidence, $U_k^\circ \equiv U_k^\circ(S)$, is obtained from a lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $L_k^\circ \equiv L_k^\circ(S)$, by replacing $1-\alpha$ by α . This completes the proof.

Theorem 8. Let $X_1 \leq \dots \leq X_r$ be the first r ordered observations of a random variable $X(=\exp X)$ from a sample of size n from a two-parameter Weibull distribution defined by the probability density function (22). Then an upper statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{U}_k^\circ \equiv \underline{U}_k^\circ(\underline{S})$, on future outcomes of the k th order statistic $Y_k(=\exp Y_k)$ from a set of m future ordered observations $Y_1 \leq \dots \leq Y_m$ also from the distribution (22), which satisfies (7) is given by

$$\underline{U}_k^\circ = \eta_{\underline{U}_k^\circ}^{1/\hat{\delta}} \hat{\beta} = \exp U_k^\circ, \quad (62)$$

where $\eta_{\underline{U}_k^\circ}$ is a tolerance factor determined by

$$\eta_{\underline{U}_k^\circ} = \arg \left(\sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{\mathfrak{g}(\mathbf{z}^{(r)})} \int_0^\infty \frac{v_2^{r-2} \prod_{i=1}^r z_i^{v_2}}{\left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} + [\eta_{\underline{U}_k^\circ}]^{v_2} (m-l+j) \right)^r} dv_2 = \alpha \right). \quad (63)$$

Proof. The upper statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{U}_k^\circ \equiv \underline{U}_k^\circ(\underline{S})$, is obtained from a lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k^\circ \equiv \underline{L}_k^\circ(\underline{S})$, by replacing $1-\alpha$ by α . This completes the proof.

Inference 4 (for the tolerance factors $\eta_{\underline{U}_k^\circ}$ and $\eta_{U_k^\circ}$). It follows from (60) and (62) that

$$\eta_{\underline{U}_k^\circ} = \left(\frac{\underline{U}_k^\circ}{\hat{\beta}} \right)^{\hat{\delta}} = \left(\frac{\exp U_k^\circ}{\exp \hat{\theta}_1} \right)^{1/\hat{\theta}_2} = \exp \left(\frac{U_k^\circ - \hat{\theta}_1}{\hat{\theta}_2} \right) = \eta_{U_k^\circ}. \quad (64)$$

NUMERICAL EXAMPLES

Numerical Example 1

For the Weibull case, Lawless (1973) discusses an example with 10 items, which are put on test simultaneously; the life test is terminated at the time of the fifth failure, whence $n=10$, $r=5$, in our notation here; $\underline{X}_1=50.5$ hours, $\underline{X}_2=71.3$, $\underline{X}_3=84.6$, $\underline{X}_4=98.7$, $\underline{X}_5=103.8$; the maximum likelihood estimates of δ and β are, respectively,

$$\hat{\delta} = \left[\left(\sum_{i=1}^r \underline{x}_i^{\hat{\delta}} \ln \underline{x}_i + (n-r) \underline{x}_r^{\hat{\delta}} \ln \underline{x}_r \right) \left(\sum_{i=1}^r \underline{x}_i^{\hat{\delta}} + (n-r) \underline{x}_r^{\hat{\delta}} \right)^{-1} - \frac{1}{r} \sum_{i=1}^r \ln \underline{x}_i \right]^{-1} = 4.199, \quad (65)$$

$$\hat{\beta} = \left(r^{-1} \left(\sum_{i=1}^r \underline{x}_i^{\hat{\delta}} + (n-r) \underline{x}_r^{\hat{\delta}} \right) \right)^{1/\hat{\delta}} = 114.2796. \quad (66)$$

Based on these data, a lower 90% prediction limit (in terms of this paper, a lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, where $\alpha=0.1$) is to be constructed for the minimum of 40 independently, identically distributed lifetimes. Lawless (1973) reports a conditional lower 90% prediction limit of 8.8 hours for this example. Based on a simulation of 50 000 samples, the lower prediction limit obtained by Mee and Kushary (1994) is 8.73 hours.

Lower statistical tolerance limit with expected $(1-\alpha)$ -confidence. Taking $1-\alpha = 0.9$ and $k=1$, with $n=10$, $r=5$ and $m=40$, we have from (57) that the lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k^\circ \equiv \underline{L}_k^\circ(S)$, on the minimum (Y_1) of independent lifetimes in a group of $m=40$ components which are to be put into service, is given by

$$\underline{L}_k^\circ = \eta_{\underline{L}_k^\circ}^{1/\hat{\delta}} \hat{\beta} = \exp L_k^\circ = \exp \left(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k^\circ} \right) = 8.7941146, \quad (67)$$

where, as it follows from (58) and (59), the tolerance factor $\eta_{\underline{L}_k^\circ}$ is given by

$$\eta_{\underline{L}_k^\circ} = \arg \left(\frac{1}{\mathcal{G}(\mathbf{z}^{(r)})} \int_0^\infty \frac{v_2^{r-2} \prod_{i=1}^r z_i^{v_2}}{\left(\sum_{i=1}^r z_i^{v_2} + (n-r) z_r^{v_2} + \left[\eta_{\underline{L}_k^\circ} \right]^{v_2} m \right)^r} dv_2 = 1 - \alpha \right) = \eta_{\underline{L}_k^\circ} = 2.105 / 10^5. \quad (68)$$

Statistical inference 1. It is easy to see that the conditional lower 90% prediction limit of 8.8 hours of Lawless (1973) on the minimum (Y_1) of independent lifetimes in a group of $m=40$ components, which are to be put into service, and the lower statistical tolerance limit of 8.7941146 hours with expected 0.9-confidence, which is obtained in this paper by using the proposed technique, are practically the same.

Lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence. In the above case (if $\gamma=0.9$), it follows from (43) that the lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k \equiv \underline{L}_k(S)$, on the minimum (Y_1) of independent lifetimes in a group of $m=40$ components which are to be put into service, is given by

$$\underline{L}_k = \eta_{\underline{L}_k}^{1/\hat{\delta}} \hat{\beta} = \exp L_k = \exp \left(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k} \right) = 3.7, \quad (69)$$

where, as it follows from (44) and (45), the tolerance factor $\eta_{\underline{L}_k}$ is given by

$$\eta_{L_k} = \arg \left(\int_0^\infty \int_0^\infty \frac{\ln(1-q_{1-\gamma})^{-1} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)}{[\eta_{L_k}]^{v_2}} \frac{1}{\Gamma(r)} w^{r-1} \exp(-w) \frac{1}{g(\mathbf{z}^{(r)})} v_2^{r-2} \prod_{i=1}^r z_i^{v_2} \left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} \right)^{-r} dw dv_2 = 1 - \alpha \right) = \eta_{L_k} = 5.5451 / 10^7, \quad (70)$$

$q_{1-\gamma}=0.002631$ is a quantile of the beta distribution satisfying (38).

Statistical inference 2. Thus, the manufacturer has 90% assurance that no failures will occur in the proportion $\gamma=0.9$ or more of the population of $m=40$ components, which are to be put into service, before $L_k=3.7$ hours.

Others have computed approximate 90% prediction limits for a single future lifetime for this example. Fertig, Meyer and Mann (1980) computed a lower prediction limit of 56.98 hours, using best linear invariant estimators and a Monte Carlo estimated percentile. Engelhardt and Bain (1982) proposed two approximations; for this example they obtained 56.8 (via a procedure requiring iterative solution of a nonlinear equation) and 59.1 (via a simpler approximation). Based on a simulation of 50 000 samples, the 90% lower prediction limit obtained by Mee and Kushary (1994) for a single future observation is 56.6 hours.

Lower statistical tolerance limit with expected $(1-\alpha)$ -confidence. Taking $1-\alpha=0.9$ and $k=m=1$, with $n=10$ and $r=5$, we have from (57) that the lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k^\circ \equiv \underline{L}_k^\circ(\underline{S})$, on a single future observation is given by

$$\underline{L}_k^\circ = \eta_{\underline{L}_k^\circ}^{1/\delta} \hat{\beta} = \exp L_k^\circ = \exp(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k^\circ}) = 56.641, \quad (71)$$

where, as it follows from (58) and (59), the tolerance factor $\eta_{\underline{L}_k^\circ}$ is given by

$$\eta_{\underline{L}_k^\circ} = \arg \left(\frac{1}{g(\mathbf{z}^{(r)})} \int_0^\infty \frac{v_2^{r-2} \prod_{i=1}^r z_i^{v_2}}{\left(\sum_{i=1}^r z_i^{v_2} + (n-r)z_r^{v_2} + [\eta_{\underline{L}_k^\circ}]^{v_2} \right)^r} dv_2 = 1 - \alpha \right) = \eta_{\underline{L}_k^\circ} = 0.052479. \quad (72)$$

Statistical inference 3. Based on a simulation of 50 000 samples, the 90% lower prediction limit obtained by Mee and Kushary (1994) for a single future lifetime is 56.6 hours, which is slightly smaller than 56.641 hours (see (71)). Engelhardt and Bain (1982) proposed the first approximate lower prediction limit of 56.8 hours and Fertig, Meyer and Mann (1980) computed the lower prediction limit of 56.98 hours for a single future lifetime, which are slightly larger than 56.641 hours (see (71)). The second approximate lower prediction limit of 59.1 hours proposed by Engelhardt and Bain (1982) for a single future lifetime is larger than 56.641 hours (see (71)).

Numerical Example 2

Consider the following results given by Lieblein and Zelen (1956) of test of endurance, in millions of revolutions, of $n=23$ ball bearings: 17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.48, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40. The maximum likelihood estimates of δ and β are, respectively,

$$\hat{\delta} = \left[\left(\sum_{i=1}^n x_i^{\hat{\delta}} \ln x_i \right) \left(\sum_{i=1}^n x_i^{\hat{\delta}} \right)^{-1} - \frac{1}{n} \sum_{i=1}^n \ln x_i \right]^{-1} = 2.102, \quad (73)$$

$$\hat{\beta} = \left(n^{-1} \left(\sum_{i=1}^n x_i^{\hat{\delta}} \right) \right)^{1/\hat{\delta}} = 81.878. \quad (74)$$

Numerous authors have used these data as illustrative of a sample from a two-parameter Weibull distribution. Using 20 000 simulated samples of size $n=23$, Mee and Kushary (1994) obtained a 90% lower prediction limit for the fifth failure out one hundred ball bearings equal to 10.11 million of revolutions, which is slightly smaller than the two approximate prediction limits 10.27 and 10.59 reported by Engelhardt and Bain (1982).

Lower statistical tolerance limit with expected $(1-\alpha)$ -confidence. Taking $1-\alpha = 0.9$ and $k=5$, with $r=n=23$ and $m=100$, we have from (57) that the lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k^{\circ} \equiv \underline{L}_k^{\circ}(\underline{S})$, for a fifth failure out one hundred ball bearings is given by

$$\underline{L}_k^{\circ} = \eta_{\underline{L}_k^{\circ}}^{1/\hat{\delta}} \hat{\beta} = \exp L_k^{\circ} = \exp \left(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k^{\circ}} \right) = 10.35206, \quad (75)$$

where, as it follows from (58) and (59), the tolerance factor $\eta_{\underline{L}_k^{\circ}}$ is given by

$$\eta_{\underline{L}_k^{\circ}} = \arg \left(\sum_{l=0}^{k-1} \binom{m}{l} \sum_{j=0}^l \binom{l}{j} (-1)^j \frac{1}{\mathfrak{g}(\mathbf{z}^{(n)})} \int_0^{\infty} \frac{v_2^{n-2} \prod_{i=1}^n z_i^{v_2}}{\left(\sum_{i=1}^n z_i^{v_2} + [\eta_{\underline{L}_k^{\circ}}]^{v_2} (m-l+j) \right)^n} dv_2 = 1-\alpha \right) = \eta_{\underline{L}_k^{\circ}} = 0.0129452. \quad (76)$$

Statistical inference 4. Thus, it follows from (75) that the lower statistical tolerance limit with expected confidence of 0.9 ($\underline{L}_k^{\circ} = 10.35206$ million of revolutions) is between the two approximate prediction limits 10.27 and 10.59 reported by Engelhardt and Bain (1982).

Lower statistical tolerance limit with expected $(1-\alpha)$ -confidence. Taking $1-\alpha = 0.9$ and $k=1$, with $r=n=23$ and $m=100$, we have from (57) that the lower statistical tolerance limit with expected $(1-\alpha)$ -confidence, $\underline{L}_k^{\circ} \equiv \underline{L}_k^{\circ}(\underline{S})$, for a first failure out one hundred ball bearings, is given by

$$\underline{L}_k^\circ = \eta_{\underline{L}_k^\circ}^{1/\hat{\delta}} \hat{\beta} = \exp L_k^\circ = \exp(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k^\circ}) = 2.083, \quad (77)$$

where, as it follows from (58) and (59), the tolerance factor $\eta_{\underline{L}_k^\circ}$ is given by

$$\eta_{\underline{L}_k^\circ} = \arg \left(\frac{1}{g(\mathbf{z}^{(n)})} \int_0^\infty \frac{v_2^{n-2} \prod_{i=1}^n z_i^{v_2}}{\left(\sum_{i=1}^n z_i^{v_2} + [\eta_{\underline{L}_k^\circ}]^{v_2} \right)^n} dv_2 = 1 - \alpha \right) = \eta_{\underline{L}_k^\circ} = 0.00044503. \quad (78)$$

Statistical inference 5. Lawless (1973) obtained for this example (via conditional approach in terms of a extreme-value (Gumbel) distribution) the lower 90% prediction limit of 2.08 million of revolutions for a first failure out one hundred ball bearings, which is slightly smaller than 2.083 (see (77)).

Numerical Example 3

Consider the data in an example discussed by Mann and Saunders (1969). They regard the data coming from the Weibull distribution as the results of full scale fatigue tests on a particular type of component. The data are for a complete sample of size $n=3$, with observations $X_1=45.952$, $X_2=54.143$, and $X_3=65.440$, results being expressed here in number of thousands of cycles. On the basis of these data it is wished to obtain the lower statistical tolerance limit with expected $(1-\alpha)$ -confidence for the minimum (Y_1) of independent lifetimes in a group of $m=500$ components which are to be put into service.

Lower statistical tolerance limit with expected $(1-\alpha)$ confidence. The maximum likelihood estimates of the unknown parameters δ and β , computed on the basis of (X_1, X_2, X_3) , are $\hat{\delta} = 7.726$ and $\hat{\beta} = 58.706$, respectively. Taking $1-\alpha = 0.8$ and $k=1$, with $r=n=3$ and $m=500$, we have from (57) that the lower statistical tolerance limit with expected $(1-\alpha)$ confidence, $\underline{L}_k^\circ \equiv \underline{L}_k^\circ(\underline{S})$, on the minimum (Y_1) of independent lifetimes in a group of $m=500$ components which are to be put into service, is given by

$$\underline{L}_k^\circ = \eta_{\underline{L}_k^\circ}^{1/\hat{\delta}} \hat{\beta} = \exp L_k^\circ = \exp(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k^\circ}) = 5.527411, \quad (79)$$

where, as it follows from (58) and (59), the tolerance factor $\eta_{\underline{L}_k^\circ}$ is given by

$$\eta_{\underline{L}_k^\circ} = \arg \left(\frac{1}{g(\mathbf{z}^{(n)})} \int_0^\infty \frac{v_2^{n-2} \prod_{i=1}^n z_i^{v_2}}{\left(\sum_{i=1}^n z_i^{v_2} + [\eta_{\underline{L}_k^\circ}]^{v_2} \right)^n} dv_2 = 1 - \alpha \right) = \eta_{\underline{L}_k^\circ} = 1.18 / 10^8. \quad (80)$$

Statistical inference 6. Lawless (1973) obtained for this example (via conditional approach in terms of the extreme-value (Gumbel) distribution) the lower 80% prediction limit of 5.623, which is slightly larger than 5.527411 (see (79)). The resulting lower 80% prediction limit of Mee and Kushary (1994) for this example (obtained via simulation of 100 000 samples) was 5.225, which is slightly smaller than

5.527411 (see (79)). The Mann and Saunders (1969) result for this example was only 0.766. All results are expressed here in the number of thousands of cycles.

Lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence. Taking $\gamma=0.8$, $1-\alpha=0.8$ and $k=1$, with $r=n=3$ and $m=500$, we have from (43) that a lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence, $L_k \equiv L_k(\mathcal{S})$, on the minimum (Y_1) of independent lifetimes in a group of $m=500$ components which are to be put into service, is

$$\underline{L}_k = \eta_{\underline{L}_k}^{1/\delta} \hat{\beta} = \exp L_k = \exp(\hat{\theta}_1 + \hat{\theta}_2 \ln \eta_{\underline{L}_k}) = 4.082282, \quad (81)$$

where, as it follows from (44) and (45), the tolerance factor $\eta_{\underline{L}_k}$ is given by

$$\eta_{\underline{L}_k} = \arg \left(\int_0^\infty \int_0^\infty \frac{\ln(1-q_{1-\gamma})^{-1} \sum_{i=1}^n Z_i^{v_2}}{[\eta_{\underline{L}_k}]^{v_2}} \frac{1}{\Gamma(n)} w^{n-1} \exp(-w) \frac{1}{g(\mathbf{z}^{(n)})} v_2^{n-2} \prod_{i=1}^n z_i^{v_2} \left(\sum_{i=1}^n z_i^{v_2} \right)^{-n} dw dv_2 = 1 - \alpha \right) = \eta_{\underline{L}_k} \\ = 1.135/10^9, \quad (82)$$

$q_{1-\gamma}$ is a quantile of the beta distribution satisfying (38).

Statistical inference 7. Thus, the manufacturer has 80% assurance that no failures will occur in the proportion $\gamma=0.8$ or more of the population of $m=500$ components, which are to be put into service, before $\underline{L}_k=4.082282$ thousands of cycles.

FUTURE RESEARCH DIRECTIONS

Predictive inferences (via tolerance limits) for future outcomes on the basis of the past and present knowledge represent a fundamental problem of statistics, arising in many contexts and producing varied solutions. In this article, new-sample prediction (via tolerance limits) based on a previous sample is considered (i.e., when for predicting the future outcomes in a new sample there are available the observed data only from a previous sample). It is interesting to consider within-sample prediction (via tolerance limits) based on the early data from a current experiment (i.e., when for predicting the future outcomes in a sample there are available the early data only from that sample), and new-within-sample prediction (via tolerance limits) based on both the early data from that sample and the data from a previous sample (i.e., when for predicting the future outcomes in a new sample there are available both the early data from that sample and the data from a previous sample), where it is assumed that only the functional form of the underlying distributions is specified, but some or all of its parameters are unspecified.

Applications of Exact Statistical Tolerance Limits on Future Random Quantities for Prognostics and Health Management of Complex Systems

Prognostics and Health Management (PHM) is a set of means, approaches, methods and tools that allows monitoring and tracking the health state of a system in order to detect, diagnose and predict its failures. Topics of interest include, but are not limited to:

- (1) Prognostics and Health Management in Cyber-Physical Systems;
- (2) Fault detection, diagnostics and prognostics (data-driven, model-based and hybrid methods);
- (3) Remaining Useful Life (RUL) computation and prediction;
- (4) Post-diagnostics and post-prognostics decision;
- (5) Condition monitoring and sensors placement and optimization;
- (6) Feature extraction and selection, Health Indicators construction;
- (7) Modelling and simulation of interdependent failure mechanisms;
- (8) Advanced computation and simulation methods;
- (9) Industrial applications.

Prognostics and Health Management (PHM) aims at development and publishing original scientific contributions and industrial applications dealing with the topics covered by PHM in the areas of condition monitoring, fault detection, fault diagnostics, fault prognostics and decision support. The results obtained are then exploited to take appropriate decisions to increase the system's availability, reliability and security while reducing its maintenance costs.

Development of Novel Technologies of Prognostics and Health Management (PHM) for l -out-of- m Systems on the Basis of the Exact Statistical Tolerance Limits

Many technical systems or subsystems have l -out-of- m structure. These so-called l -out of- m systems consist of m components of the same kind. The entire system is working if at least l of its m components are operating. It fails if $m - l + 1$ or more components fail. Hence, a l -out-of- m system breaks down at the time of the $(m - l + 1)$ th component failure. Since all components start working at the same time, this approach leads to a kind of redundancy called active redundancy of $m - l$ components. Important particular cases of l -out-of- m systems are parallel and series systems corresponding to $l = 1$ and $l = m$, respectively. Practical examples of l -out-of- m systems are, e.g., an aircraft with four engines which will not crash if at least two out of its four engines remain functioning, or a satellite which will have enough power to send signals if not more than four out of its ten batteries are discharged.

Prediction problems come up naturally in several real life situations. For example, they can be broadly classified under two categories: (i) the variable to be predicted comes from the same experiment or sample so that it may be correlated with the observed data, (ii) it comes from an independent future experiment. Both of these situations do arise in the context of reliability and lifetesting. Formally the problems in the first category are known as one-sample (or within-sample) problems and those in the latter constitute two-sample (or new-sample) problems as well as multi-sample problems.

Suppose a system consists of m components and fails whenever l of these components fail. Such a system is referred to as l -out-of- m system. Suppose our observations consist of the first k failure times, and the goal is to predict the failure time of the system. Assuming that the components' life lengths are identically and independently distributed, we have a prediction problem involving a Type II censored sample, and it falls into category (i). Formally, the problem is to predict Y_l having observed $(Y_1 \leq \dots \leq Y_k)$, $k < l$. A popular choice for the lifelength of the components is the exponential distribution. Thus, one can think of a point predictor or an interval predictor for the lifelength of the system as one-sample (or within-sample) problem.

Consider the third situation where a manufacturer of certain equipment is interested in setting up a warranty for the equipment in a lot being sent out to the market. Using the information based on a small

sample, possibly censored, the goal is to predict and set a lower prediction limit for the weakest item in a future sample. Typical assumption here is that the two samples are independent. This falls into category (ii), and we call it a two-sample (or new-sample) problem. One may be interested in the lifelength of the l -th weakest item and the average lifelength simultaneously. Then our focus will be on a prediction region for the two variables of interest.

While several researches have concentrated on the above problems there are some papers dealing with special prediction situations involving several independent random samples. These belong to category (ii) and can be taken into account.

In reliability theory, the lifetime of a l -out-of- m system is usually described by the $(m - l + 1)$ th order statistic Y_{m-l+1} from the sample $Y_1 \leq \dots \leq Y_m$, where the random variable Y_i represents the lifetime or failure time of the i th component of the system, $1 \leq i \leq m$. In the conventional modelling of these structures, the component lifetimes are supposed to be independent and identically distributed random variables. *Translating this approach back into the technical sphere, it reflects the assumption that the failure of any component does not affect the remaining ones. However, the supposition that the breakdown of some component will have no impact on the system parts at work will generally not be fulfilled in practice. In some systems, a component failure will more or less strongly influence the remaining parts of the system. For example, the breakdown of an aircraft's engine will increase the load put on the remaining engines such their lifetimes tend to be shorter. Thus, a more flexible model, which is therefore more applicable to practical situations, must take some dependence among the system components into account.*

Prognostics and Health Management (PHM) offers several benefits for predictive maintenance. It predicts the future behavior of l -out-of- m system as well as its Remaining Useful Life (RUL). This RUL is used to plan the maintenance operation to avoid the failure, the stop time and optimize the cost of the maintenance and failure. However, with the development of the industry the assets are nowadays distributed, this is why the PHM needs to be developed using the new Information Technologies. In this paper, we propose a PHM solution based on the exact statistical prediction and tolerance limits on future random quantities (for example, future order statistics coming from Log-Location-Scale Distributions). These statistical limits will be constructed via the Pivotal Quantity Averaging Approach (PQAA), which represents the conceptually simple, efficient and useful method for constructing exact statistical prediction and tolerance limits on future outcomes under parametric uncertainty of underlying models. Nechval et al. (2014b, 2014c) discuss some of the problems described in this section.

Development of Novel Technologies of Prognostics and Health Management (PHM) for Fatigued Structures on the Basis of the Exact Statistical Tolerance Limits

From an engineering standpoint the fatigue life of a fatigued structure consists of two periods: (i) crack initiation period, which starts with the first load cycle and ends when a technically detectable crack is presented, and (ii) crack propagation period, which starts with a technically detectable crack and ends when the remaining cross section can no longer withstand the loads applied and fails statically. The main aim of this paper is to present more accurate innovative stochastic fatigue model for adaptive planning inspections of fatigued structures in damage tolerance situations via observations of crack growth process during a crack propagation period. A new crack growth equation is based on this model. It is attractively simple and easy to apply in practice for effective in-service inspection planning (with decreasing intervals between sequential inspections as alternative to constant intervals often used in practice for convenience in operation). During the period of crack propagation (when the damage tolerance situation is used), the

proposed crack growth equation, based on the innovative model, allows one to construct more accurate and effective reliability-based inspection strategy in this case.

Prognostics and Health Management (PHM) is a technology to enhance the effective reliability and availability of a fatigued structure in its life cycle conditions by detection of current and approaching failures and by providing for mitigation of the structure risks. Prognostics is the real-time enhancement of reliability and availability and the prediction of the remaining useful life of the structure by assessing the extent of deviation or degradation of the monitored parameters of the structure from its expected normal operating conditions. Prognostics can yield an advance warning of impending failure in a structure, thereby enabling more efficient and effective maintenance and corrective actions. Prognostics help in preventing catastrophic failures and can reduce unscheduled maintenance expenses. The outputs of a prognostic assessment of a structure are the failure risk, time to failure, remaining useful life, and a prognostic distance within which maintenance and repair actions can be planned, in order to minimize their impact on system availability. Health Management is the process of using diagnostic and prognostic information to intelligently manage the use and maintenance of a system. The ultimate result is increased effective reliability, availability, and safety with reduced logistics and support costs.

Prognostics and Health Management (PHM) can be used to solve the service problems of the following important engineering structures: (1) Transportation Systems and Vehicles – aircraft, space vehicles, trains, ships; (2) Civil Structures – bridges, dams, tunnels; (3) Power Generation – nuclear, fossil fuel and hydroelectric plants; (4) High-Value Manufactured Products – launch systems, satellites, semiconductor and electronic equipment; (5) Industrial Equipment – oil and gas exploration, production and processing equipment, chemical process facilities, pulp and paper. . Nechval et al. (2015, 2018b) discuss some of the problems described in this section.

Development of Novel Technologies of Prognostics and Health Management (PHM) for a New Product in Systems of Lifetime Testing on the Basis of the Exact Statistical Tolerance Limits

A new product lifetime testing represents the problem that can be stated as follows. A new product is submitted for lifetime testing. The product will be accepted if a random sample of n items shows r or fewer failures in performance testing. We want to know whether to stop the test before it is completed if the results of the early observations are unfavourable. A suitable stopping decision saves the cost of the waiting time for completion. On the other hand, an incorrect stopping decision causes an unnecessary design change and a complete rerun of the test. It is assumed that the redesign would improve the product to such an extent that it would definitely be accepted in a new lifetime testing. The paper presents a stopping rule based on the statistical estimation of total costs involved in the decision to continue beyond an early failure. Sampling is both expensive and time consuming. Hence, there are situations where it is more efficient to take samples sequentially, as opposed to all at one time, and to define a stopping rule to terminate the sampling process.

At the planning stage of a statistical investigation the question of sample size (n) is critical. For such an important issue, there is a surprisingly small amount of published literature. Engineers who conduct reliability tests need to choose the sample size when designing a test plan. The model parameters and quantiles are the typical quantities of interest. The large-sample procedure relies on the property that the distribution of the t -like quantities is close to the standard normal in large samples. To estimate these quantities the maximum likelihood method is often used. The large-sample procedure to obtain the sample size relies on the property that the distribution of the above quantities is close to standard

normal in large samples. The normal approximation is only first order accurate in general. When sample size is not large enough or when there is censoring, the normal approximation is not an accurate way to obtain the confidence intervals. Thus sample size determined by such procedure is dubious. Therefore, it may be considered the problem of constructing a test which minimizes the maximum expected sample size under some constraints. Stopping rules in sample testing can save substantial time and resources, when the case is clear-cut.

Prognostics and Health Management (PHM) is essential in guaranteeing the safe, efficient, and correct operation of complex of detection, isolation and identification of faults; and prognosis, which consists of prediction of the remaining useful life (RUL) of components, subsystems, or systems, constitutes system health monitoring. PHM aims to provide users with an integrated view of the health state of equipment or overall system. Nechval et al. (2009) discusses some of the problems described in this section.

CONCLUSION

In this article, we construct the following one-sided statistical tolerance limits: i) one-sided statistical tolerance limit that covers at least $100\gamma\%$ of the measurements with expected $100(1-\alpha)\%$ confidence, ii) one-sided statistical tolerance limit determined so that the expected proportion of the measurements covered by this limit is $(1-\alpha)$. Tolerance limits have important role in application of statistical methods in technical practice, especially in statistical quality control. Inherent in every phase of industrial quality control is the problem of comparing some quality characteristic or measurement of a finished product against given specifications. Sometimes the specifications, or *tolerance limits*, are so stated by the customer or by design engineer that any appreciable departure will make the product unusable. There remains, however, the problem of producing the part so that an acceptably high proportion of units will fall within tolerance limits specified for the given quality characteristic. Also, if a product is made without prior specifications, or if modifications are made, it is desirable to know within what limits the process can hold a quality characteristics a reasonably high percentage of the time. We thus speak of natural tolerance limits; that is, we let the process establish its own limits which, according the experience, can be met in actual practice.

The new analytical technique proposed in this article represents the conceptually simple, efficient and useful method for constructing exact statistical tolerance limits on future outcomes under parametric uncertainty of underlying models. It does not in need to make any assumption concerning the statistical equation for the tolerance limit. This technique, using the experimental complete or type II censored data, is based on the idea of invariant embedding of a sample statistic in the underlying model to construct pivotal quantities and to eliminate the unknown parameters from the problem via pivotal quantity averaging. In this case, the exact statistical tolerance limits (under parametric uncertainty of underlying models) on future outcomes (say, order statistics) associated with sampling from corresponding distributions can be found easily and quickly making tables, simulation and special computer programs unnecessary.

The analytical methodology described in this paper is illustrated for the two-parameter Weibull and extreme-value distributions. Applications to other log-location-scale distributions could follow directly.

Finally, we give the three illustrative numerical examples, where the exact statistical tolerance limits with expected $(1-\alpha)$ -confidence, obtained in this paper in terms of the two-parameter Weibull or extreme-value distribution, are compared with the known results that are reported in the literature and were obtained by using the following: 1) tables, 2) simulation, 3) Monte Carlo estimated percentiles, 4)

special computer programs, 5) approximation, 6) transformation of the two-parameter Weibull distribution to the extreme-value distribution, etc.

Although the details of the problems addressed in the article can vary significantly from one industry to another, the focus is always on making more accurate decisions, rather than manually using guesses and intuitions, but rather from a scientific point of view using models and technologies implemented with disciplined processes and systems.

The methodologies described here can be extended in several different directions to solve various problems arising in practice.

REFERENCES

- Dunsmore, J. R. (1978). Some approximations for tolerance factors for the two parameter exponential distribution. *Technometrics*, 1978(20), 317–318. doi:10.1080/00401706.1978.10489677
- Engelhardt, M., & Bain, L. J. (1978). Tolerance limits and confidence limits on reliability for the two-parameter exponential distribution. *Technometrics*, 20(1), 37–39. doi:10.1080/00401706.1978.10489615
- Engelhardt, M., & Bain, L. J. (1982). On prediction limits for samples from a Weibull or extreme-value distribution. *Technometrics*, 24(2), 147–150. doi:10.1080/00401706.1982.10487738
- Fertig, K. W., Meyer, M. E., & Mann, N. R. (1980). On constructing prediction intervals from a Weibull or extreme-value distribution. *Technometrics*, 22(4), 567–573. doi:10.1080/00401706.1980.10486206
- Guenther, W. C. (1972). Tolerance intervals for univariate distributions. *Naval Research Logistics Quarterly*, 19(2), 309–333. doi:10.1002/nav.3800190208
- Guenther, W. C., Patil, S. A., & Uppuluri, V. R. R. (1976). One-sided β -content tolerance factors for the two parameter exponential distribution. *Technometrics*, 18, 333–340.
- Gumbel, E. J. (1958). *Statistics of Extreme*. Columbia University Press. doi:10.7312/gumb92958
- Hahn, G. J., & Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*. John Wiley & Sons. doi:10.1002/9780470316771
- Lawless, J. F. (1973). On the estimation of safe life when the underlying life distribution is Weibull. *Technometrics*, 15(4), 857–865. doi:10.1080/00401706.1973.10489118
- Lieblein, J., & Zelen, M. (1956). Statistical investigation of the fatigue life of deep-groove ball bearing. *Journal of Research of the National Bureau of Standards*, 47(5), 273–316. doi:10.6028/jres.057.033
- Mann, N. R., & Saunders, S. C. (1969). On evaluation of warranty assurance when life has a Weibull distribution. *Biometrika*, 56(3), 615–625. doi:10.1093/biomet/56.3.615
- Mee, R. W., & Kushary, D. (1994). Prediction limits for the Weibull distribution utilizing simulation. *Computational Statistics & Data Analysis*, 17(3), 327–336. doi:10.1016/0167-9473(92)00072-Y
- Nechval, K. N., & Nechval, N. A. (1999). Constructing lower simultaneous prediction limits on observations in future samples from the past data. *Computers & Industrial Engineering*, 137(1-2), 133–136. doi:10.1016/S0360-8352(99)00039-X

Nechval, K. N., Nechval, N. A., Berzins, G., & Purgailis, M. (2009). Optimal statistical decisions in a new product lifetime testing. *Proceedings of the Fourth International Conference on Maintenance and Facility Management*, 135–140.

Nechval, N. A., Berzins, G., Balina, S., Steinbuka, I., & Nechval, K. N. (2017). Constructing unbiased prediction limits on future outcomes under parametric uncertainty of underlying models via pivotal quantity averaging approach. *Automatic Control and Computer Sciences (AC & CS)*, 51, 331–346.

Nechval, N. A., Berzins, G., & Nechval, K. N. (2018b). Intelligent planning reliability-based inspections of fatigued structures for the crack initiation period in the Weibull case under parametric uncertainty. *Automatic Control and Computer Sciences*, 52(3), 184–197. doi:10.3103/S0146411618030094

Nechval, N. A., Berzins, G., Nechval, K. N., & Krasts, J. (2019). A New Technique of Intelligent Constructing Unbiased Prediction Limits on Future Order Statistics Coming from an Inverse Gaussian Distribution under Parametric Uncertainty. *Automatic Control and Computer Sciences (AC & CS)*, 53, 223–235.

Nechval, N. A., & Nechval, K. N. (2014a). A new approach to constructing simultaneous prediction limits on future outcomes under parametric uncertainty of underlying models. In A. Sio-long, C. A. Hoi-Shou, K. Hideki, & X. Li (Eds.), *IAENG Transactions on Engineering Sciences* (pp. 1–13). Taylor & Francis Group. doi:10.1201/b16763-2

Nechval, N. A., & Nechval, K. N. (2015). Improved planning in-service inspections of fatigued aircraft structures under parametric uncertainty of underlying lifetime models. In *Numerical Methods for Reliability and Safety Assessment: Multiscale and Multiphysics Systems*. Springer International Publishing. doi:10.1007/978-3-319-07167-1_25

Nechval, N. A., & Nechval, K. N. (2016a). Tolerance limits on order statistics in future samples coming from the two-parameter exponential distribution. *American Journal of Theoretical and Applied Statistics*, 5, 1–6. doi:10.11648/j.ajtas.s.2016050201.11

Nechval, N. A., Nechval, K. N., & Berzins, G. (2014c). A new technique for intelligent constructing exact γ -content tolerance limits with expected $(1 - \alpha)$ -confidence on future outcomes in the Weibull case using complete or Type II censored data. *Automatic Control and Computer Sciences (AC & CS)*, 52, 476–488.

Nechval, N. A., Nechval, K. N., & Berzins, G. (2018a). A new technique for constructing exact tolerance limits on future outcomes under parametric uncertainty. In M. Ram & J. P. Davim (Eds.), *Advanced Mathematical Techniques in Engineering Sciences* (pp. 203–226). Taylor & Francis Group. doi:10.1201/b22440-11

Nechval, N. A., Nechval, K. N., Danovich, V., & Berzins, G. (2014b). Predictive inferences for future order statistics coming from an inverse Gaussian distribution. *Proceedings of The World Congress on Engineering*, 888–893.

Nechval, N. A., Nechval, K. N., Prisyazhnyuk, S. P., & Strelchonok, V. F. (2016b). Tolerance limits on order statistics in future samples coming from the Pareto distribution. *Automatic Control and Computer Sciences (AC & CS)*, 50, 423–431.

Nechval, N. A., Nechval, K. N., & Strelchonok, V. F. (2016c). A new approach to constructing tolerance limits on order statistics in future samples coming from a normal distribution. *Advances in Image and Video Processing*, 4(4), 47–61. doi:10.14738/aivp.42.2014

Nechval, N. A., & Vasermanis, E. K. (2004). *Improved Decisions in Statistics*. Riga: Izglitibas soli.

Patel, J. K. (1986). Tolerance limits: A review. *Communications in Statistics. Theory and Methods*, 15(9), 2719–2762. doi:10.1080/03610928608829278

KEY TERMS AND DEFINITIONS

Extreme-Value Distribution: The so-called first asymptotic distribution of extreme values, hereafter referred to simply as the extreme-value distribution, which is extensively used in a number of areas as a lifetime distribution and sometimes referred to as the Gumbel distribution.

Future Outcome: The value of the k th order statistic, $1 \leq k \leq m$, in a future random sample of m ordered observations $Y_1 \leq \dots \leq Y_m$.

Parametric Uncertainty: It is assumed that only the functional form of the underlying distributions is specified, but some or all of its parameters are unspecified. In this article, it is assumed that both parameters of the underlying distributions are unknown.

Statistical γ -Content Tolerance Limit with Expected $(1-\alpha)$ -Confidence: An upper statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence is determined so that with the given expected confidence level $1-\alpha$, a specified proportion γ or more of the population will fall below the limit. A lower statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence satisfies similar conditions.

Statistical Tolerance Limit with Expected $(1-\alpha)$ -Confidence: An upper statistical tolerance limit with expected $(1-\alpha)$ -confidence is determined so that the expected proportion of the population failing below the limit is $(1-\alpha)$. A lower statistical tolerance limit with expected $(1-\alpha)$ -confidence satisfies similar conditions.

Two-Parameter Weibull Distribution: This two-parameter distribution is one of the most widely used life distributions in reliability analysis. It is very flexible, and can, through an appropriate choice of parameters, model many types of failure rate behaviors. The distribution has wide applications in diverse disciplines.

Type II Censored Data: In this case, we only observe, say, the r smallest observations in a random sample of n items.

The Structure of DNA Taking Into Account the Higher Dimension of Its Components

Gennadiy Vladimirovich Zhizhin

 <https://orcid.org/0000-0002-4933-7590>

Skolkovo, Russia

INTRODUCTION

Deoxyribonucleic acid (DNA), as a chemical substance, it was isolated by Johann Friedrich Misher in 1869 from the remains of cells contained in the pus. He singled out a substance that includes nitrogen and phosphorus. When Misher determined that this substance has acid properties, the substance was called nucleic acid (Dahm, 2005). Gradually it was proved that it was DNA, and not proteins, as previously thought, and which is the carrier of genetic information. One of the first decisive proofs was the experiments of Oswald Avery, Colin MacLeod and McLean McCarthy (1944) on the transformation of bacteria.

The structure of the double helix DNA it was proposed by Francis Crick and James Watson in 1953 on the base of the X-ray structural data obtained by Maurice Wilkins and Rosalind Franklin and the “Chargaff rules” according to which in each DNA molecule the strict relationships connecting the quantity of nitrogenous bases of different (Watson, & Crick, 1953a, b). For outstanding contributions to this discovery, Francis Crick, James Watson and Maurice Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine. Deoxyribonucleic acid (DNA) is a biopolymer, the monomer of which is the nucleotide (Albert, et al., 2002; Butler, 2005). Each nucleotide consists of a phosphoric acid residue attached to sugar deoxyribose, to which one of the four nitrogen bases is attached also. The bases that make up the nucleotides are divided into two groups: purines (adenine [A] and guanine [G]) and pyrimidines (cytosine [C] and thymine [T]) are formed by combined five - and six - membered heterocycles.

They managed to show that the DNA isolated from the pneumococci corresponds to the so-called transformation (the acquisition of pathogenic properties by a harmless culture as result of the addition of dead pathogenic bacteria to it). The experiment of American scientists Alfred Hershey and Martha Chase (Hershey - Chase experiment, 1952) with radioactively labeled proteins and bacteriophage DNA showed that only the phage nucleic acid is transmitted to the infected cell, and the new generation of phage contains the same proteins and nucleic acid, as the initial phage (Hershey & Chase, 1952). Deciphering the structure of DNA (1953) has become one of the turning points in the history of biology.

In 1986, Frank-Kamenetskiy in Moscow showed how a double-stranded DNA folds into a so-called H - shape, composed not of two but three strands of DNA (Frank – Kamenetskiy, 1986, 1988). Deoxyribonucleic acid (DNA) is a biopolymer, the monomer of which is the nucleotide (Albert et al., 2002; Butler, 2005).

Nucleotides are long polynucleotide chains covalently linked. These chains in the overwhelming majority of cases (except for some viruses possessing single - stranded DNA genomes) are combined pairwise by means of hydrogen bonds into a secondary structure, called the double helix (Watson & Crick, 1953 a, b; Berg, Tymoczko & Stryer, 2002). Each base on one of the chains is connected to one definite base on the second chain. This specific binding is called complementary. Purines are complementary to pyrimidines (that is, they are capable of forming hydrogen bonds with them): adenine forms bonds

DOI: 10.4018/978-1-7998-3479-3.ch050

only with thymine, and cytosine - with guanine. In a double helix, chains are also linked by hydrophobic interactions and stacking, which do not depend on the DNA base sequence (Ponnuswamy & Gromiha, 1994). Complementarity of the double helix means that the information contained in one chain is also contained in another chain. Different base pairs form a different number of hydrogen bonds. In the future, the existence of nucleic acids differing in the length of the period and shape with rotation of the spiral both to the right and to the left was experimentally established (Ha, et al., 2005; Cantor, & Schimmel, 1980; Frank – Kamenetskiy, 2010).

Watson and Crick postulated the spiral form of the DNA molecule, but they did not discuss the reasons for the formation of such a DNA molecule. Until now there have been no works explaining the existence of a spiral in the DNA molecule.

In the book (Zhizhin, 2017), the molecules of practically all the elements of the periodic system were studied in detail and it was shown that many of them, including magnesium and calcium, form compounds of higher dimension (see also Zhizhin, 2015 a, b; Zhizhin, & Diudea, 2016; Zhizhin, Khalaj, & Diudea, 2016). In the works (Zhizhin, 2016 a, b), the structures of biomolecules, which also form compounds of higher dimension, are also studied. When analyzing the metric space of nucleic acid structures, one should use the representations of their constituents in the form of polytopes of higher dimension. This primarily applies to the phosphoric acid residues and the *D* - ribose molecules, since nitrogenous bases have a dimensionality of 2 and do not have a significant effect on the geometric structure of nucleic acids.

The determination of the geometric forms of chemical compounds according to the Euler-Poincaré (Poincare, 1895) relation translates classical stereochemistry in three-dimensional space into sulfur chemistry in a space of higher dimension. Here new discoveries are possible in the laws of chemical compounds. Consideration of the geometry of molecules using spaces of higher dimension makes it possible to explain the facts observed in reality, which previously had no clear explanations. In the book (Zhizhin, 2017), for example, showed that the rotation of the plane of light polarization in tartaric acid is due to the difference in the shapes of the molecules of *D* - tartaric acid and *L* - tartaric acid, which is clearly visible in their images in five - dimensional space, since their dimension is five. Pasteur (Pasteur, 1960) also tried to explain the rotation of the polarization plane of light in liquid tartaric acid by the difference in the structures of the crystals of *D* - tartaric acid and *L* - tartaric acid. This book received an explanation for the fact that there is no branching in the chain of glucose molecules, if they are bound in β - glycoside linkage. If these molecules are bound in α - glucoside linkage, branching of the chain is possible. It is shown that, because of the higher dimensionality of glucose molecules in the case of a β - glycoside linkage, the centers of possible branching of the chain are blocked by glucose molecules. In this time, with α - glycoside linkage the possible branching points of the chain remain free.

The concept of the geometric correspondence of the structure of substances in the formation of a chemical compound is essential for the form of the compound. However, the complementarity of DNA spirals when connecting spirals can not be the reason for the formation of these spirals. The geometric shape of the constituent parts of the spiral itself must be important here. They are phosphoric acid residues and *D* - ribose molecules. This paper is devoted to the investigation of this problem. In this case it is necessary to take into account the dimension of these parts, which, as shown in (Zhizhin, 2017), is higher than three. It should be noted, that there is no contradiction in the representation of the structure of DNA as a spiral in a three - dimensional space with the highest dimension of its components. The point is that according to the geometry of Riemann (Riemann, 1854), a space of higher dimension has a boundary. Therefore, a space of a smaller dimension can be located outside this space.

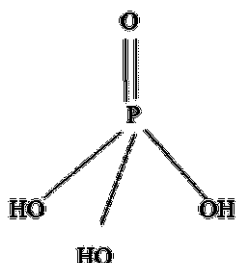
The higher dimensionality of molecules can be considered as an example of Big Data. The concept of Big Data has become widespread in computer science in recent times (Kosrapohr, 2003; Hughes,

2011). The representation of molecules in the form of polytopes of higher dimension means the handling of Big Data as geometric images of a set of points in a space of higher dimension. A striking example of the use of the concept of Big Data in medicine is a work of J.A. Rodger (Rodger, 2015) in which, using the intellectual analysis of Big Data, a program has been developed to improve the prognosis of survival rates for traumatic brain injury and other injuries on several ships. The totality of injury data for each vessel can be considered as Big Data in view of the many parameters characterizing trauma.

THE DIMENSION OF PHOSPHORIC ACID AND ITS RESIDUE

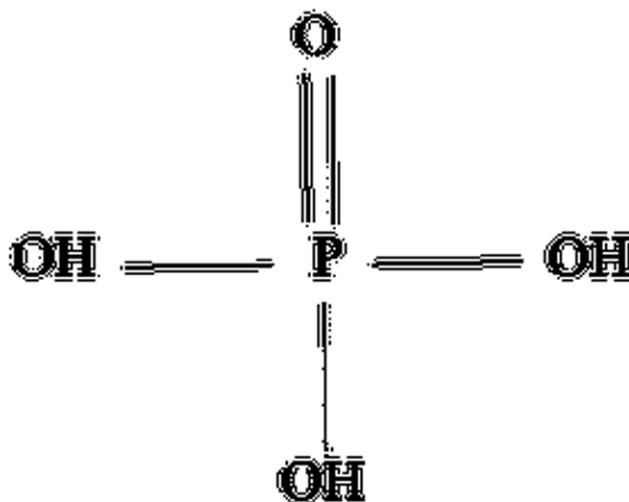
The phosphoric acid H_3PO_4 has the following building (see, for example, Karapetians & Drakin, 1994) (Figure 1).

Figure 1. Special structure of phosphoric acid



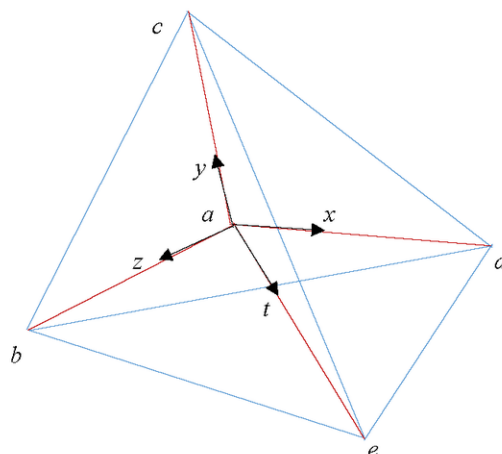
On the Figure 2 is presented Fisher's formula of phosphoric acid.

Figure 2. Fisher's formula of phosphoric acid



It is obvious that, used in the DNA molecule, the phosphoric acid residue obtained by detaching one hydrogen atom from it has the building also. In this case, instead of one of the hydroxyl groups attached to the phosphorus atom, there will be a negatively charged oxygen atom. To determine the functional topological dimension of phosphoric acid and its residue, it is necessary to connect the vertices, atoms and functional groups entering into these compounds (Zhizhin, 2016b, 2017). Then the structure of phosphoric acid and its residue is represented in the form of a polytope in Figure 3.

Figure 3. The polytope of the phosphoric acid and its residue



In Figure 3 in vertices a dispose atom phosphoric; in vertices c dispose atom of oxygen with double bond; in vertices b, d, e dispose atoms of oxygen with free bond.

Calculation of the functional topological dimension by the Euler-Poincaré equation (Poincaré, 1895) does not require specifying the exact lengths of polytope edges, and for a given number of the vertices does not depend on the groups of atoms in these vertices (Zhizhin, 2016b). In Figure 3, the edges corresponding to the covalent chemical bonds are marked in red, and the edges that carry only the geometric function, delineating the closed geometric figure, are marked in blue.

It follows from Figure 3 that the polytope of phosphoric acid has five vertices (a, b, c, d, e), ten edges ($ac, ab, ad, ae, cd, de, be, ce, cb, bd$), ten plane triangular faces ($cbe, cda, cea, dab, db e, dce, bae, dae, bce, dcb$), and five three - dimensional faces in the form of tetrahedron ($cbae, caed, cbad, abde, cebd$). These values of elements of different dimensions we must substitute into the Euler - Poincare equation

$$\sum_{i=0}^{n-1} f_i (-1)^i = 1 - (-1)^n, \quad (1)$$

where f_i is the number of elements with dimension i in polytope with dimension n . In this case there are $f_0=5, f_1=10, f_2=10, f_3=5$.

Substituting these values into equation (1), can to obtain

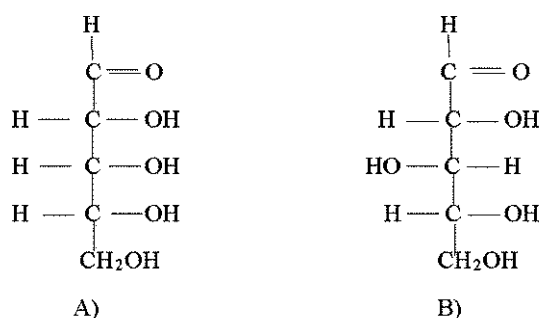
$$5 - 10 + 10 - 5 = 0.$$

Thus, equation (1) is satisfied for $n = 4$. This proves that the polytope of phosphoric acid and its residue has a dimension of four. It is worth emphasizing that the polytope of phosphoric acid and its residue is topologically equivalent to a convex polytope of the simplex type with dimension four, since these polytopes completely coincide with the values of the numbers of elements of different dimensions (Zhizhin, 2014b). The origin of the four - dimensional space can be placed in one of the vertices with sending coordinates along the edges connecting this vertex to other vertices (see Figure 3).

THE DIMENSION OF THE DEOXYRIBOSE

The deoxyribose (or the *D* – ribose) is molecule of carbohydrate with five atoms carbon. There are two enantiomer forms of this molecule. Its images in the form of Fisher's formula (Metzler, 1980) are shown in Figure 4 A, B.

Figure 4. A, B. Two enantiomer forms of the molecule *D* – ribose (formulas of Fisher's)



In the DNA molecule, *D* - ribose as a component enters with a closed carbon chain. The closed chain of the molecule in Figure 4 is formed due to the rupture of the double bond between oxygen and carbon (considering it the first in the chain), the liberation of the water molecule, and the addition of this oxygen atom to the penultimate (fourth) carbon atom in the carbon chain. The Fischer formula or formula Haworth and their modifications (Metzler, 1980; Lehninger, 1982) may not reflect the spatial structure of the *D* – ribose molecule. For this target the constructs described in the form of convex polytopes with boundary elements which form boundary complex (Grunbaum, 1967). Only when such a representation will be to determine the dimension of these molecules (Zhizhin, 2016 b; 2017). Figure 5 shows the result of the closure of a chain of four carbon atoms through an oxygen atom correspondent Figure 4 A. Chemical bonds are marked as before with red color.

It is easy to see that, in addition to the form of the *D* - ribose molecule in Figure 5 (a form of A), there is another enantiomer form of this molecule. It connected by a symmetry transformation with respect to a straight line passing through an oxygen atom separating the projection of the form of A in half (we shall assume that it is a B shape). Form B of the *D* - ribose molecule is shown in Figure 6.

In order to give the *D* - ribose molecule a kind of polytope, it is necessary to connect each vertex (every atom or functional group, with all other vertices) to the ribs, then can to get a polytope of the simplex type, the dimension of which must be determined.

Figure 5. The molecule *D* – ribose with closed a chain of carbon atoms (the form A)

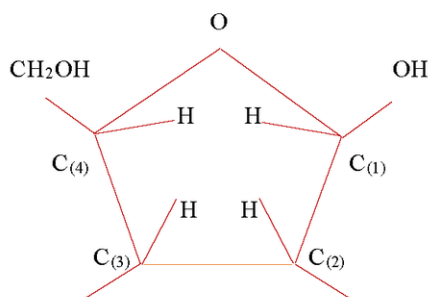


Figure 6. The molecule *D* – ribose with closed a chain of carbon atoms (the form B)

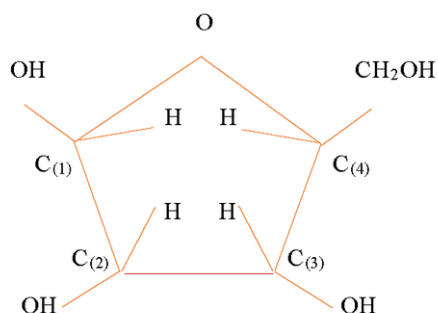
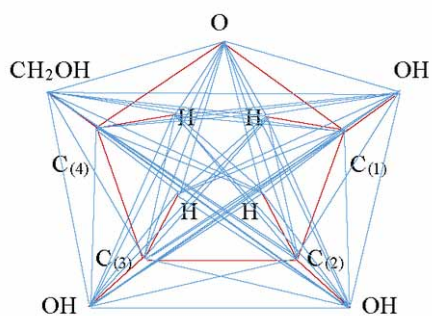


Figure 7 shows an image of a *D* - ribose molecule (A) with a closed carbon chain in the form of a polytope.

Figure 7. The molecule *D* – ribose (A) in form of the polytope



In contrast to the previous works of the author (Zhizhin, 2016 b; 2017), all hydroxyl groups are transferred to the periphery of the image of the molecule. This is more convenient for considering possible chemical reactions of *D* - ribose with other compounds.

The polytope at Figure 7 has 13 vertices in which there are one oxygen atom, four hydrogen atoms, three hydroxyl groups, four carbon atoms, compound CH_2OH . It is known that the dimension of a simplex is one less than the number of its vertices (Grünbaum, 1967). Therefore, the dimension of the polytope in Figure 7, if the Euler - Poincaré equation (1) holds for it, equals $n=f_0 - 1=12$. The simplex is a simplicial polytope, and all its faces therefore are simplexes. Since all the vertices of a simplex are related to each other, the number of edges is determined by the number of combinations of the number of vertices of two. In this case, the number of edges of the polytope in Figure 7 is $f_1 = C_{n+1}^2 = 78$.

The number of two - dimensional faces (triangles) of this polytope is $f_2 = C_{n+1}^3 = 286$.

The number of three - dimensional faces (tetrahedrons) of this polytope is $f_3 = C_{n+1}^4 = 715$.

The number of four - dimensional faces (4 - simplexes) of this polytope is $f_4 = C_{n+1}^5 = 1287$.

The number of five - dimensional faces (5 - simplexes) of this polytope is $f_5 = C_{n+1}^6 = 1716$.

The number of six - dimensional faces (6 - simplexes) of this polytope is $f_6 = C_{n+1}^7 = 1716$.

The number of seven - dimensional faces (7 - simplexes) of this polytope is $f_7 = C_{n+1}^8 = 1287$.

The number of eight - dimensional faces (8 - simplexes) of this polytope is $f_8 = C_{n+1}^9 = 715$.

The number of nine - dimensional faces (9 - simplexes) of this polytope is $f_9 = C_{n+1}^{10} = 286$.

The number of ten - dimensional faces (10 - simplexes) of this polytope is $f_{10} = C_{n+1}^{11} = 78$.

The number of eleven - dimensional faces (11 - simplexes) of this polytope is $f_{11} = C_{n+1}^{12} = 13$.

Substituting the values f_i , ($0 \leq i \leq 11$) in the Euler - Poincaré equation (1), can to see that it holds for $n = 12$ for polytope on Figure 7

$$13 - 78 + 286 - 715 + 1287 - 1716 + 1716 - 1287 + 715 - 286 + 78 - 13 = 0.$$

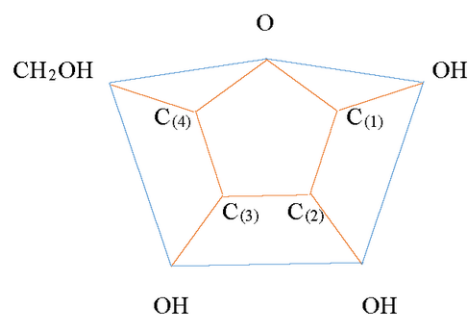
This finally proves that the polytope D - ribose (A) has dimension 12 and a simplex type. Obviously, the polytope D - ribose (B) has the same dimension and type.

THE STRUCTURE OF THE DEOXYRIBOSE NUCLEIC ACID

The DNA is a chain of D - ribose molecules between which are the phosphoric acid residues that connect these molecules. Let us construct this circuit, taking into account the concept of repulsion of electron pairs (Gillespie, 1972; Gillespie & Hargittai, 1991). According to this concept, tetrahedral coordination of atoms and functional groups in these compounds is carried out around the phosphorus atom in the phosphoric acid residue, and in the carbon atoms in the D - ribose molecule also. It is this concept that leads to the higher dimensionality of the remainder of phosphoric acid and the molecule of D - ribose. When a phosphoric acid residue is added to the D - ribose molecule from the hydroxyl group of the phosphoric acid residue and the hydroxyl group of the D - ribose molecule, a water molecule that leaves the compounds forms and an oxygen atom that binds the phosphoric acid residue and the D - ribose molecule. The concept of repulsion of electron pairs also operates here, according to which two valence electrons of the oxygen atom maximally repel each other, forming electron pairs with electrons of the phosphorus atom and the carbon atom. Therefore, taking into account that tetrahedral coordination is carried out around the atoms of phosphorus and carbon already, it can assume that the valence bonds of the oxygen atom are located linearly with respect to the oxygen atom. When the second D - ribose molecule is added to the phosphoric acid residue, the second hydroxyl group of the phosphoric acid

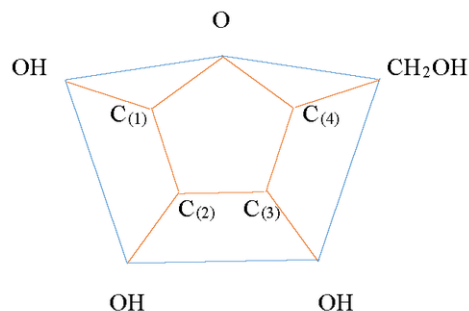
residue with the compound CH_2OH of this *D* - ribose molecule also form molecules of water that leaves the compound and an oxygen atom connecting the phosphorus atom and the functional group CH_2 of the *D* - ribose molecule. By virtue of the concept of repulsion of electron pairs, the linear arrangement of the valence bonds of the oxygen atom can be considered here also. In addition, as the first step within the framework of the notions of the functional dimension of molecules (Zhizhin, 2016 b), can shall assume the linear arrangement of the valence bonds also in the neighborhood of the functional group CH_2 of the *D* - ribose molecule. One will take into account this when constructing a chain of DNA molecules and phosphoric acid residues. To construct the sequence of *D* - ribose molecules and the phosphoric acid residue in the chain can will use a simplified image of the *D* - ribose molecule of dimension 12. Let us leave only the outer contour and valence bonds necessary for interaction with the environment from the *D* - ribose molecule. (Figure 4 B). At the same time, one will not depict the hydrogen atoms inside the contour of carbon atoms. Also one approximately to consider that the lengths of chemical bonds carbon - carbon, carbon - oxygen close to each other. The image of the *D* - ribose molecule of the enantiomer form A under these conditions takes the form of Figure 8.

Figure 8. The simplified image of the enantiomer A of *D* – ribose molecule



The corresponding image of the *D* - ribose molecule of the enantiomer form B is shown in Figure 9.

Figure 9. The simplified image of the enantiomer B of *D* – ribose molecule

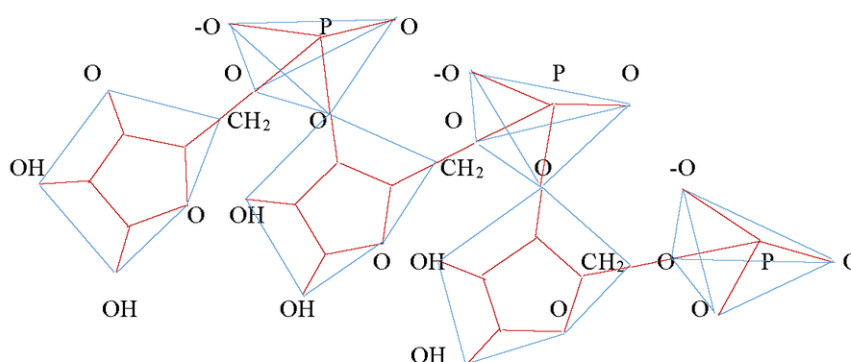


The absence of edges and atoms in Figure 8 and Figure 9 in comparison with the Figure 7 does not mean that they really are not. This is only the conventionality of the image.

It should be noted that under the assumed projection conditions, the *D* - ribose of molecules on a two - dimensional plane in which a closed carbon ring is located has a number of properties that are convenient for further constructions. The enclosed carbon ring is a regular pentagon with an internal angle at the vertices $\pi - \frac{2\pi}{5} = 108^\circ$. The angles in the three trapezoids formed by the valence bonds and the outer contour, because of the symmetry and parallelism of the bases, are 54 degrees and 126 degrees. In two obtuse - angled isosceles triangles formed by valence bonds and the outer contour, the angles are 126 degrees and 27 degrees.

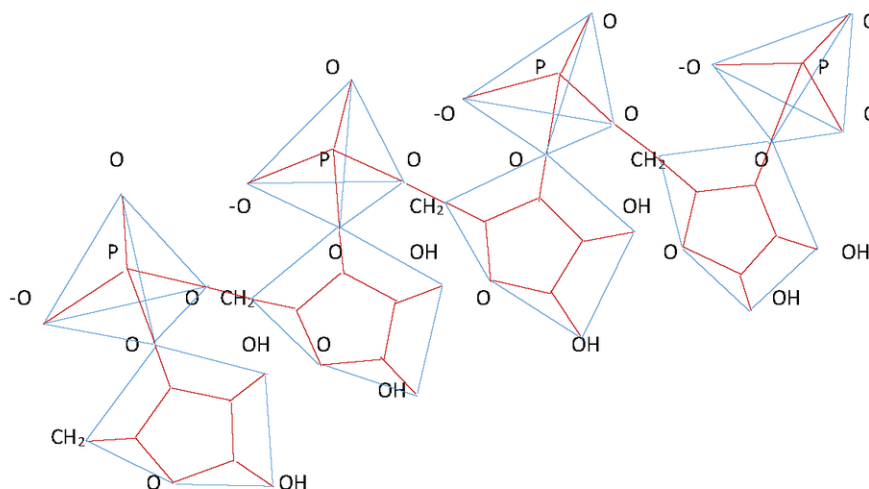
Now, in a projection onto a two - dimensional plane, one shall depict a sequence of the simplified images of *D* - ribose molecules and phosphoric acid residues, taking into account the separation of water molecules at the junction of phosphoric acid residues and *D* - ribose molecules and the linear arrangement of the valence bonds in vicinity of these compounds. On Figure 10 presented the sequence for *D* – ribose molecules in form A.

Figure 10. Structure of the DNA molecule with D – ribose molecules in form A



On Figure 11 presented the sequence for *D* – ribose molecules in form B.

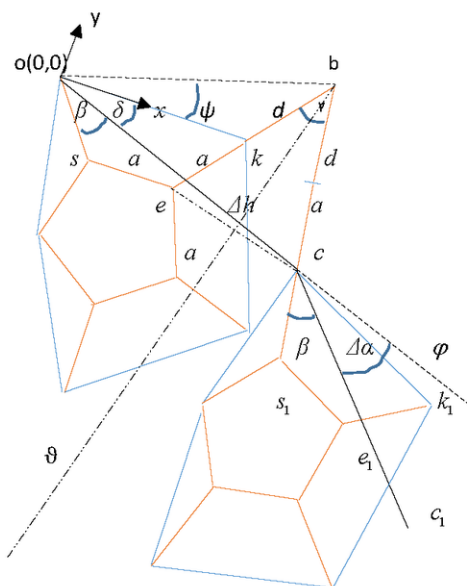
Figure 11. Structure of the DNA molecule with D – ribose molecules in form B



It follows from Figures 10 and 11 that the *D* - ribose molecules in both chains are not linearly distributed. Any sequence of corresponding points of *D* - ribose molecules in chains forms a curved line with rotation to the right side. However, in these sequences there is a fundamental difference. It is a consequence of the use in these two DNA structures of two enantiomeric molecules of *D* - ribose: the *D* - ribose molecule in form *A* in the DNA chain in Figure 10, the *D* - ribose molecule in the *B* form in the DNA chain in Figure 11. This difference consists in the fact, that in the DNA chain in Figure 10 the oxygen atom in the carbon ring of the *D* - ribose molecule is the extreme right in each *D* - ribose molecule. And in the DNA chain in Figure 11 the oxygen atom in the carbon ring of the *D* - ribose molecule is the extreme left in the carbon ring of the molecule *D* - ribose. It can be argued that the DNA structures in Figures 10 and 11 are enantiomeric DNA structures, which are a consequence of the presence of two enantiomeric structures of *D* - ribose molecules in them. The existence of two enantiomeric DNA structures was not noted before, since details of the arrangement of atoms in the DNA structures were not considered. Denote the structure of the DNA in Figure 10 of the DNA of the *EA*, and the structure of the DNA in Figure 11 of the DNA of the *EB*.

The projection of the DNA molecules depicted in Figures 10 and 11 can be described mathematically. Let us turn to a detailed picture of the connection of two nuclei (two *D* - ribose molecules of the form *B*) in Figure 11 (Figure 12).

Figure 12. Two nuclei of the DNA structure of the *EB*



The vertex *b* in this figure corresponds to the phosphorus atom, and the two edges emanating from it correspond to the projections of two valence bonds of the atom phosphorus with two oxygen atoms. They are at some angle γ on the projection. This angle characterizes the location of the residual phosphoric acid in space with respect to the *D* - ribose molecule *B* at the site of their connection (vertex *k*). It can be shown that the angle γ is linearly related to the angle $\Delta\alpha$ between two neighboring nuclei in the spiral chain of nuclei. Indeed, in Figure 12, the smallest distance between two corresponding points of two neighboring *D* - ribose molecules *B* in the chain Δh corresponds to the segment *OC*. The small-

est angle $\Delta\alpha$ characterizing the helical surface corresponds to the angle between the straight line that is the extension of the OC segment and the next segment CC_1 between two molecules of D - ribose, originating from the vertex C . The angle $\Delta\alpha = \angle ocb - \beta$, where $\angle ocb = \angle s_1cp$, β is the angle between segment OC and edge OS in the initial molecule D - ribose. The same angle is repeated in the next molecule of D - ribose. The angle $\beta = 54^\circ - \delta$ where 54° is the angle between edge OS and edge OK . The angle $\angle obc = \angle obk + \gamma$, but $\angle obk = \pi - \angle okb - \psi$, $\angle okb = 126^\circ$ by construction, therefore $\angle obk = 54^\circ - \psi$.

On the other hand $\angle ocb = \pi - \angle cob - \angle obc = \pi - (\delta + \psi) - \angle obk - \gamma = 126^\circ - \delta - \gamma$.

Therefore $\Delta\alpha = 126^\circ - \delta - \gamma - (54^\circ - \delta) = \frac{\pi}{5} - \gamma$.

An amazing result was obtained: the rotation angle between two molecules of D - ribose in the chain depends linearly only on the angle between the projections of the valence bonds in the phosphoric acid residue, that is, the orientation of the phosphoric acid residue in the space with respect to the D - ribose molecule

$$\Delta\alpha = \frac{\pi}{5} - \gamma. \quad (2)$$

Concordantly with model of Watson and Crick (Watson & Crick, 1953 a, b) in period of nuclei chain included 10 nuclei. In this chain the angle $\Delta\alpha = 36^\circ$. Therefore in this case, according to (2), the angle γ is also equal to 36 degrees. If in the period of nuclei chain include 12 nuclei (see Ha, et. el, 2005), so $\Delta\alpha = 30^\circ$, and $\gamma = 42^\circ$.

It is possible to calculate the minimum distance between the corresponding points of neighboring D - ribose molecules in the nucleic chain Δh . It follows from Figure 12 that $\Delta h = \sqrt{y_c^2 + x_c^2}$.

Values y_c, x_c are from the system

$$4(d+a)^2 \cos^2 \frac{\pi - \gamma}{2} = (y_e - y_c)^2 + (x_e - x_c)^2, (d+a)^2 = (y_b - y_c)^2 + (x_b - x_c)^2, \quad (3)$$

where $x_e = a(1 + \cos 54^\circ)$, $y_e = -a \sin 54^\circ$, $x_b = a(1 + 2 \cos 54^\circ) + d \sin 36^\circ$, $y_b = d \cos 36^\circ$.

The first equation of system (3) can be obtained from the equality of the square of the length of the segment ec to the sum of the squares of the differences of the coordinates of its vertices, on the one hand. On the other hand, the square of the length of the segment ec is the square of the doubled product of the cosine of the angle at the base of the isosceles triangle ecb by the length of the force side. The second equation of system (3) is obtained from the equality of the square of the side length cb to square of the sum of the segments a and d , on the one hand, and to the sum of the squares of the differences of the coordinates of its vertices, on the other hand. Solving system (3) for given a, d , one can find the value of Δh . To find the solution, the squares of the coordinate differences in the right-hand sides of the equations of system (3) are opened and the second equation is subtracted from the first. The linear connection of the coordinates of the vertex c is obtained. The equation of this connection is substituted into the second equation of system (3) and a quadratic equation is obtained for one of the coordinates of

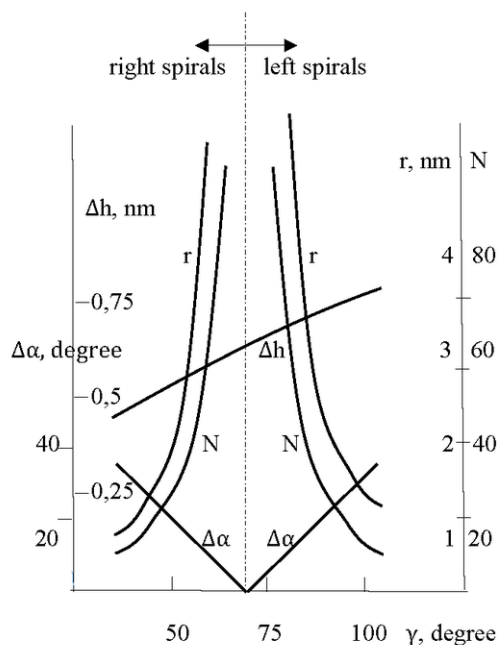
vertex c . From two roots of the equation, the root corresponding to Figure 12 is chosen. Thus, for each value of the angle γ for given the values of a, d are the coordinates of the vertex c .

It is easy to verify that by mirroring the image on Figure 12 with respect to the bisector $b\theta$ of the angle γ passing through the vertex b , one can obtain an image of two coupled nucleotides corresponding to another enantiomeric form of DNA (Figure 10). In this way, the calculation of the parameters of the DNA structure according to the system (3) corresponds directly to two enantiomeric forms of the DNA structure. In the beginning, in order to find out the properties of the solutions of the system (3) corresponding to possible DNA structures, it can will numerically study these solutions, arbitrarily varying the value of the angle γ , which is the defining parameter of this system. Note that in this study it confine ourselves to analyzing one chain of a DNA molecule, although, as is known (Watson & Crick,

Table 1. The characteristic dimensions of the right and left DNA helices

Γ degree	$\Delta\alpha$ degree	N	Δh Nm	R Nm
36	36	10	0,4677	0,7447
42	30	12	0,501	0,9575
64,5	7,5	48	0,607	4,6397
72	0	∞	0,6394	∞
84	12	30	0,6815	3,2557
90	18	20	0,7	2
102	30	12	0,7337	1,4
108	36	10	0,7639	1,2163

Figure 13. Graphs of the change in the distance between nucleotides Δh , the angle of rotation between adjacent nucleotides $\Delta\alpha$, the number of nucleotides in the period N and the radius of the spiral $r = \frac{N\Delta h}{2\pi}$, depending on the angle α in the projection between the valence bonds of the phosphorus atom.

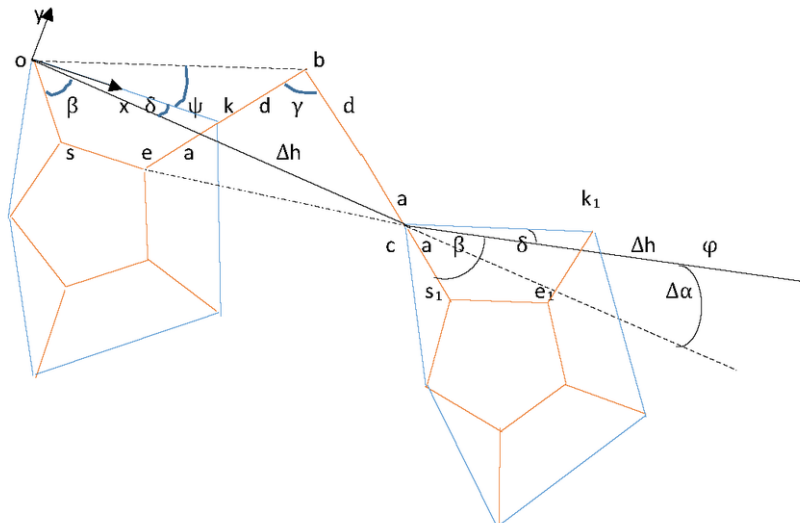


1953 a, b; Frank – Kamenetskiy, 1988, 2004, 2010), a DNA molecule consists of two or three chains of the DNA molecules connected to each other by hydrogen bonds. Such a connection of chains can be carried out in the future, relying on the obtained regularities of one chain.

Since, the lengths of chemical bonds between carbon - carbon and carbon - oxygen are ≈ 0.15 nm, can will adopt $a=0.15$ nm. The length of chemical bond phosphor – oxygen can will adopt $d = 0.18$ nm. The results of calculating the characteristics of the DNA structure from the equations of system (3) are shown in the graphs (Figure 13) and the numerical values of some parameters are given in the Table 1.

As can be seen from Figure 13 and Table 1 extension of the angle γ between the projections of the valence bonds of the phosphorus atom in the phosphoric acid residue from the angle of 36 degree to the plane of the carbon ring of the *D* - ribose molecule increases almost linearly the distance between the nucleotides in the DNA. In this case, the radius of the DNA helix increases nonlinearly asymptotically to infinity in the neighborhood of the value of the angle $\gamma = 72$ degrees. At a value γ of 72 degrees, the DNA molecule is straightened out into a straight line. With further increase in angle γ , the DNA helix begins to rotate to the left. In this range of values, increasing the angle γ also leads to an almost linear increase in the distance between nucleotides, but to a nonlinear decrease in the radius of the DNA helix. This occurs up to an angle $\gamma = 110$ degrees, at which the spiral turns into a closed ring, since there is no rise of nucleotides with the help of phosphoric acid residues. In cases where the angle is greater than 72 degrees $\Delta\alpha = \gamma - 72$. In Figure 14, as an example, an image of two adjacent nucleotides at 90 degrees is presented.

Figure 14. Two nuclei of the DNA structure of the EB for $\gamma = 90$ degrees



It can be seen from this figure that when γ is greater than 72 degrees, the angle $\Delta\alpha$ between two adjacent segments is laid to the left of the initial segment. This leads to the rotation of the spiral to the left.

Than from Table 1 can to see that for $\gamma = 36$ degree the distance between nucleotides in structure DNA $\Delta h = 0,4677$ nm. Wherein the number of nucleotides in the period is 10 and the radius of the spiral $r = 0,7447$ nm. To this value of the radius should be added the size of the phosphoric acid residue, which is along the length of the valence bond d and the angle between the valence bonds of 110 degrees $\Delta r = \sin 55^\circ d \approx 0.29$ nm. Then the radius of the DNA helix is $r + \Delta r \approx 1$ nm. The same DNA radius it was

observed in experiments (Watson & Crick, 1953 a, b) for 10 of nucleotides in the period. This form of DNA structure was designated B. From Table 1 can be seen that for $\gamma = 42$ degree the distance between nucleotides in structure DNA $\Delta h = 0,5$ nm. Wherein the number of nucleotides in the period is 12 and the radius of the spiral $r = 0,957$ nm. Then the radius of the DNA helix is $r + \Delta r \approx 1,35$ nm. These data roughly correspond to the form of the DNA structure, designated A. It is also known from experiments the existence of a DNA structure with 12 nucleotides in a period with left rotation (Ha, et al., 2005; Frank – Kamenetskiy, 2004). This form of DNA structure was designated Z. From Table 1 it follows that the value of the angle $\gamma = 102$ degrees corresponds to this form. The existence of DNA structures in the form of a closed ring has also been experimentally confirmed (Frank - Kamenetskiy, 2010). This form of DNA structure corresponds to the angle $\gamma = 110$ degrees in Table 1.

Thus, from the variety of solutions of the system (3) corresponding to the possible DNA structures, four forms of the DNA structure were experimentally confirmed. This does not mean that other forms of DNA structure that correspond to other solutions of system (3) do not exist. It should be assumed that these solutions have not yet been found in experiments, and in the future, when creating the appropriate conditions, they will be found. The received decisions should be considered also at consideration of forms of structures of DNA consisting of three connected spirals (form H) (Frank - Kamenetskiy, 1953 a, b). In addition, must not forget that each solution of the system (3) corresponds simultaneously to two enantiomeric forms of the DNA structure.

FUTURE RESEARCH DIRECTIONS

Further development of the accepted geometrical model of DNA structure with participation of elements of higher dimension - a residue of phosphoric acid and a molecule of *D* - ribose. In particular, to abolish some simplifying assumptions about the geometry of the chemical bond between the phosphoric acid residue and the *D* - ribose molecule.

To investigate the enantiomeric structures of DNA *EA* and *EB* by combining two or three DNA helices of a certain geometry into one compound using bases.

To investigate the effect of the chemical and physiological conditions on the geometric chemical bonds of the phosphoric acid residue and the *D* - ribose molecule.

Identify the features of the structure of RNA within the accepted geometric model.

To study the features of replication and translation for the protein synthesis in ribosomes, taking into account the developed geometric model of DNA.

Obtaining experimental data on the distribution of atoms and functional groups over the ribosome space would allow us to calculate the dimensionality of the ribosome space and to clarify the nature of the effect of this dimension on the processes of protein synthesis with given hereditary information.

CONCLUSION

The currently known forms of the helical structure of DNA molecules *A*, *B*, *Z*, and DNA molecules in the form of a closed ring were experimentally found in the works of Watson, J.D., Crick, F.H.C., Ha, S.C., Frank - Kamenetskiy, M. and other scientists. However, the reasons for the formation of such a different form of DNA molecules have not been clarified. In this paper is proved that the main reason for the formation of DNA molecules of different shapes is the different geometric realization of the

chemical bond between the phosphoric acid residue and the *D* - ribose molecule, as objects of higher dimensionality. In other words, the residues of phosphoric acid can geometrically act as a bridge, connecting the molecules of *D* - ribose with each other. It is shown that the angle γ between the projections of the valence bonds of the phosphorus atom in the phosphoric acid residue on the plane of the carbon ring in the *D* - ribose molecule is a parameter characterizing the various geometric forms of chemical bonding of phosphoric acid residues with *D* - ribose molecules. The angle γ characterizes the orientation of the phosphoric acid residue in the three - dimensional space. This parameter can therefore be regarded as an internal degree of freedom of the structure of the DNA molecule.

A mathematical model describing the geometric features of the chemical bond of these objects is constructed. Calculations on this model showed that the characteristics of known DNA structures, determined experimentally, and calculated from the constructed model, are in good agreement with each other, while the model does not contain any empirical parameters. Calculations show that known DNA structures are particular cases in the model and it can be considered that known examples of DNA structure forms are experimental confirmation of the correctness of the mathematical model. However, according to the mathematical model, its solutions are many other DNA structures that have not been experimentally discovered so far. One can hope that their experimental discovery is ahead.

The existence of two enantiomeric forms of DNA molecules, connected with the existence of two enantiomeric forms of *D* - ribose molecules, is established.

The resulting variety of DNA molecule structures is represented in the form of analytical dependencies, graphs and table.

They should be taken into account when analyzing the synthesis of proteins, replication and translation of DNA in the work of ribosomes (Stewart & McLachlan, 1975; Spirin & Gavrilova, 1971; Spirin, 1986).

The high dimensionality of the constituent DNA molecules unites the problems of the DNA molecule structure with the Big Data problem, which has become widespread in computer science recently in describing a variety of complex phenomena (Rodger, 2015).

More Plato suggested that objects of Nature, their quality and operating between the forces, may be the result of exposure to hidden from us the geometric structures (Yau & Nadis, 2010). In recent decades, a shining example of the opening of the geometric structure is string theory. Based on the solutions of Einstein's equations, this theory postulated that the manifold of higher dimension hidden in every point in space (Colabi -Yau manifold of dimension 6). It has vanishingly small size. In the nanoworld, as shown in (Zhizhin, 2014 a, b; Zhizhin, 2015 a, b; Zhizhin & Diudea, 2016; Zhizhin, Khalaj & Diudea, (2016); Zhizhin, 2016 a, b) is also a hidden dimension but objects (especially molecules), having these dimensions, are tangible sizes. A common basis for understanding the hidden dimensions of the objects of nature, and in string theory, explains the structure of the universe and the structure of elementary particles (Green, 2000; Gepner, 1988), and in the theory of the nanoworld is Riemann geometry (Riemann, 1854), allowing the existence of spaces of higher dimension in the space of lower dimension.

The higher dimensionality of the constituent DNA molecules, which allows to describe mathematically the structure of DNA, requires reconsidering the issues of tight packing of DNA molecules in cells, viruses and bacteria, provided that the DNA chains necessary for the preservation and transfer of genetic information are complementary.

REFERENCES

- Albert, B. (2002). *Molecular Biology of the Cell* (4th ed.). Garland Science.
- Berg, J., Tymoczko, J., & Stryer, L. (2002). *Biochemistry*. W.H. Freeman & Co Ltd.
- Butler, J. M. (2005). *Forensic DNA Typing*. Gardners Books.
- Cantor, C. R., & Schimmel, P. R. (1980). *Biophysical Chemistry. Part I. The Conformation of Biological Macromolecules*. W.H. Freeman & Co Ltd.
- Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 278(2), 274–288. doi:10.1016/j.ydbio.2004.11.028 PMID:15680349
- Gepner, D. (1988). Yukawa Couplings for Calabi -Yau String Compactification. *Nuclear Physics B*, 311(1), 191–204. doi:10.1016/0550-3213(88)90147-2
- Gillespie, R. J. (1972). *Molecular Geometry*. Van Nostrand Reinhold Company.
- Gillespie, R. J., & Hargittai, I. (1991). *The VSEPR Model of Molecular Geometry*. Allyn and Bacon.
- Green, B. (2000). *The Elegant Universe*. Vintage Books.
- Grunbaum, B. (1967). *Convex Polytopes*. Springer.
- Ha, S. C., Lowenhaupt, K., Rich, A., Kim, Y.-G., & Kim, K. K. (2005). Cristal structure of a junction between B – DNA and Z – DNA reveals two extruded bases. *Nature*, 437(7062), 1183–1186. doi:10.1038/nature04088 PMID:16237447
- Hershey, A., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36(1), 39–56. doi:10.1085/jgp.36.1.39 PMID:12981234
- Hughes, G. (2011). *How big is 'big data' in healthcare?* <http://blogs.sas.com/content/hls/2011/10/21/how-big-is-big-data-in-healthcare/>
- Kamenetskiy. (1986). DNA structure: A simple solution to the stability of the double helix? *Nature*, 324, 325.
- Kamenetskiy. (1988). H – form DNA and the hairpin – triplex model. *Nature*, 333, 214.
- Kamenetskiy. (2004). *The DNA Age*. Moscow: Publishing house KDU.
- Kamenetskiy. (2010). Queen of a living cell. From the structure of DNA to the biotechnological revolution. Moscow: Act – Press.
- Karapet'yants, M. X., & Drakin, S. I. (1994). *General and Inorganic Chemistry*. Chemistry.
- Kosrapohr. (Ed). (2003). *Managing data mining technologies in organizations: techniques and applications*. Hershey, PA: Idea Group Publishing.
- Lehninger, A. L. (1982). *Principles of Biochemistry* (Vol. 1-3). Worth Publisher, Inc.
- Metzler, D. E. (1980). *Biochemistry. The Chemical Reactions in Living Cells* (Vol. 1-3). Academic Press.

Oswald, A. T., MacLeod, C. M., & McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *The Journal of Experimental Medicine*, 79(2), 137–158. doi:10.1084/jem.79.2.137 PMID:19871359

Pasteur, L. (1960). *Selected works*. Publishing House of the Academy of Sciences of the USSR.

Poincare, A. (1895). Analysis situs. *J de e'Ecole Polytechnique.*, 1, 1–121.

Ponnuswamy, P., & Gromiha, M. (1994). On the conformational stability of oligonucleotide duplexes and tRNA molecules. *Journal of Theoretical Biology*, 169(4), 419–432. doi:10.1006/jtbi.1994.1163 PMID:7526075

Riemann, G. F. B. (1854). *On the Hypotheses Which Lie at the Foundations of Geometry*. Gottingen Observatory.

Rodger, J. A. (2015). Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. *Informatics in Medicine Unlocked*, 1, 17–26. doi:10.1016/j.imu.2016.01.002

Spirin, A.S. (1986). *Molecular biology. Structure of the ribosome and protein biosynthesis*. Moscow: High school.

Spirin, A. S., & Gavrilova, L. P. (1971). *The Ribosome*. Science.

Stewart, M., & McLachlan, A. D. (1975). Fourteen actin-binding sites on tropomyosin? *Nature*, 257(5524), 331–333. doi:10.1038/257331a0 PMID:1161036

Watson, J. D., & Crick, F. H. C. (1953a). Molecular structure of nucleic acids. *Nature*, 171(4356), 738–740. doi:10.1038/171737a0 PMID:13054693

Watson, J. D., & Crick, F. H. C. (1953b). Genetical implications of the structure of deoxyribose nucleic acid. *Nature*, 171(4361), 964–967. doi:10.1038/171964b0 PMID:13063483

Yau, S.-T., & Nadis, S. (2010). *The Shape of Inner Space. String Theory and the Geometry of the Universe's Hidden Dimensions*. Basic Books.

Zhizhin, G. V. (2014a). On the higher dimensions in nature. *Biosphere*, 6(4), 313–318.

Zhizhin, G. V. (2014b). *World 4D*. Polytechnic Service.

Zhizhin, G. V. (2015a). The dimensions of supramolecular compounds. *Biosphere*, 7(2), 149–154.

Zhizhin, G. V. (2015b). Dimensions of compounds in supramolecular chemistry. *Int. J. Chem. Modeling*, 7(2), 215–238.

Zhizhin, G. V. (2016a). From multidimensional phase spaces to multidimensionality of space in Nature. *Biosphere*, 8(3), 258–267.

Zhizhin, G. V. (2016b). The structures, topological and functional dimension of biomolecules. *J Chemoinformatics and Chemical Engineering*, 5(2), 44–58.

Zhizhin, G. V. (2017). *Chemical Compound Structures and the Higher Dimension of Molecules: emerging Research and Opportunities*. IGI Global.

Zhizhin, G. V., & Diudea, M. V. (2016). Space of nanoworld. In M. V. Putz & C. M. Marius (Eds.), *Sustainable nanosystems, development, properties, and applications* (pp. 221–245). IGI Global.

Zhizhin, G. V., Khalaj, Z., & Diudea, M. V. (2016). Geometrical and topology dimensions of the diamond. In A. R. Ashrafi & M. V. Diudea (Eds.), *Distance, symmetry and topology in carbon nanomaterials* (pp. 167–188). Springer. doi:10.1007/978-3-319-31584-3_12

KEY TERMS AND DEFINITIONS

Deoxyribonucleic Acid: A biopolymer, the monomer of which is the nucleotide.

Dimension of the Space: The number of independent parameters needed to describe the change in position of an object in space.

Nucleotide: A phosphoric acid residue attached to sugar deoxyribose, to which one of the four nitrogen bases is attached also.

Polytope: A polyhedron in the space of higher dimension.

Simplex: A convex polytope, any two vertices of which are joined by an edge.

Section 6

Data Mining and Databases

Methods and Techniques of Data Mining

6

Ana Funes

Universidad Nacional de San Luis, Argentina

Aristides Dasso

Universidad Nacional de San Luis, Argentina

INTRODUCTION

“Can machines think?”, this is the question that Turing (1950) put forward already in 1950. It is possible to consider the discussion in that article as a starting point in history of what nowadays is called Artificial Intelligence (AI), although, previous to that time, there are some works that could be considered as what is now called AI.

Already in antiquity, philosophers and thinkers in general had advanced the idea that “mechanical men” could be constructed and they would be able to mimic human beings in action and thought. So it could be said that the concept of artificial intelligence has existed for a very long time previous to 1950, the approximate date considered here as the start of what is now referred as AI.

AI has had from the 50s a fruitful history. Along the years several branches have emerged such as Machine Learning, Data Mining, Robotics, Statistical Learning, Deep Learning and others. Research on the subject and the development of tools and methods have grown significantly, specially over the past twenty years. The constant search for the causes of diseases, the improvement of automatic diagnoses methods, financial data analytics and market tendencies, among others, are only some of the innumerable applications where analysis and discovery of new patterns have fuelled the research and development of new methods, all related to Machine Learning, knowledge extraction from what is now being called Big Data, Knowledge Discovery in Databases or KDD, and Data Mining. This chapter focuses on Data Mining methods and algorithms as approaches to different learning tasks. The relationship of Data Mining with other disciplines inserted in the area of Artificial Intelligence is also analyzed, as well as a brief discussion on issues associated to the discipline and future research directions are also presented.

BACKGROUND

Theoretical discussions about the possibility of creating automatons or supposedly intelligent machines have been coming for a long time. In Ernst & Newell (1969) a research was conducted on the development of a computer program with general problem-solving capabilities. The program could solve problems like the Tower of Hanoi or prove theorems in first-order predicate calculus.

Weizenbaum, J. (1976), author of a then well-known computer program initially called ELIZA and later DOCTOR that allowed parodying the role of a Rogerian psychotherapist, was surprised, he says, for three events: (i) A number of psychiatrists thought that the program could become an automatic form of psychotherapy. (ii) How quickly and deeply people became emotionally involved with the program. (iii) The dissemination of the belief that the program showed a general solution to the problem of understanding natural language by computers. This situation reported by Weizenbaum can be compared

DOI: 10.4018/978-1-7998-3479-3.ch051

with The Turk, the well-known chess player who for decades, from the end of the 18th century until the end of the 19th century, was believed to be a machine capable of playing chess. This, and many others around the same time, can be predecessors of programs that nowadays are considered to be part of the general field of AI.

In his textbook Wiston (1977) says that “The central goals of AI are to make computers more useful and to understand the principles which make intelligence possible”.

Many authors agree that AI is an important knowledge area that is closely related to other disciplines. As Russell, S. J. & Norvig, P. (2010) express, “Artificial Intelligence (AI) is a big field ... which encompasses logic, probability, and continuous mathematics; perception, reasoning, learning, and action; and everything from microelectronic devices to robotic planetary explorers”. AI comprises not only Machine Learning but other disciplines such as Robotics, Logic Programming and Inductive Logic Programming (ILP), a field of Machine Learning and Data Mining. These different areas could conceivably be either form part of AI or have sprout from it.

AI has had different goals along its history. Initially the idea embodied in Turing’s question was paramount. Could a computer be built that acts like a human being?, hence, could it solve the problems that a human could? Several tests were developed to assess such ability. Turing (1950) put forward what is now called the Turing Test, but that he called the ‘The Imitation Game’. On the Turing Test there are different opinions. Cohen, Paul R. (2006) summarizes them saying that there are three kinds of arguments against Turing’s test, namely some say that the test is irrelevant, it should be forgotten, get rid of it, however it still stands today. There are competitions, variations on the test, etc. Another argument against the test is related to the philosophy connected to the question of whether machines can think, or what is thinking and if machines really can do it or is it just a simulacrum. Others give a methodological argument when asking the question. is the Imitation Game, –as Turing put forward– a valid methodological means of definitely determining whether a machine can think?, “They argue that the test is methodologically flawed and is based in bad philosophy, that it exposes cultural biases and naïveté about what Turing calls the “programming” required to pass the test.” Cohen, Paul R. (2006, p. 61). For more on the Turing test see Hernández-Orallo, J. (2000).

AI has greatly evolved since Turing’s paper times. That evolution has boosted the development of different branches, stimulated by practical concerns such as pattern recognition, data analysis and knowledge discovery. One of these areas is Data Mining, in which this chapter is focused.

The development of some AI fields has benefited from the existence of large volumes of data proceeding from the most diverse sources and domains. The Knowledge Discovery in Databases (KDD) process and methods of Data Mining allows for the discovery of knowledge in data that is hidden to humans, particularly when data volumes are large or even extremely large. These methods present the knowledge extracted under different ways: rules, equations, decision trees, etc., and help in answering questions such as what are the groups from a population of individuals with common characteristics?, is a client reliable?, is an e-mail spam?, etc.

Answers to these kind of questions are different from those obtained from traditional queries in On Line Transactions Processing (OLTP), where the information is not hidden neither is discovered, but it is summarized in an agreed format or report. They also differ from information proceeding from Online Analytical Processing (OLAP), which can be presented in different perspectives or aggregated in different ways and not just summarized as in OLTP, and that can even escalate to the use of big data, where OLTP fails. However, both of these methods are not capable of discovering new knowledge neither producing new patterns nor rules as the KDD process permits.

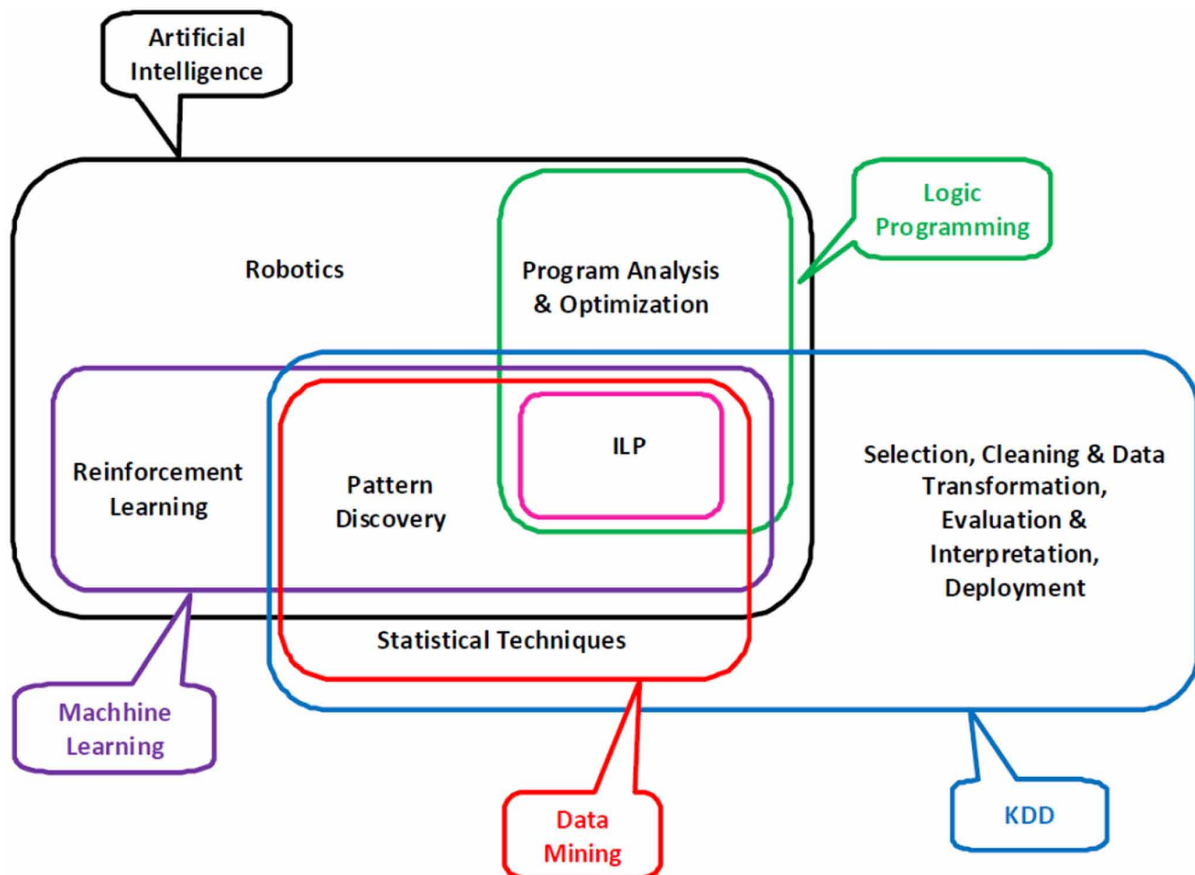
This chapter focuses on the Data Mining phase of KDD process, its tasks and methods, as well as its relation to other areas of AI such as Machine Learning, Inductive Logic Programming, Statistics, etc. A discussion on issues and challenges associated to the field is given and some future research trends are presented.

DATA MINING, MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

The notion of Data Mining is not new. Since the 60s, other terms as Data Fishing or Data Dredging have been used by statisticians to refer to the idea of finding correlations in data without a previous hypothesis as underlying causality. However, it is not until the late 80s that Data Mining became a discipline of Computer Science and scientific community adopted the term. In fact, as Witten and Frank (2011) point out, the first book on data mining appeared in 1991 (Piatetsky-Shapiro and Frawley, 1991) –a collection of papers presented at a workshop on knowledge discovery in databases in the late 1980s.

Data Mining is a branch of AI closely related to Machine Learning, where Machine Learning provides the technical basis for Data Mining (Witten and Frank, 2011). Data Mining deals with inductive learning in a practical and not theoretical way (as Machine Learning does), making use of tools provided by Machine Learning. It applies Machine Learning techniques as well as other statistical and algebraic techniques to find structural patterns hidden in data, with the main objective of describing data or making predictions from them.

Figure 1. Data Mining and other related disciplines



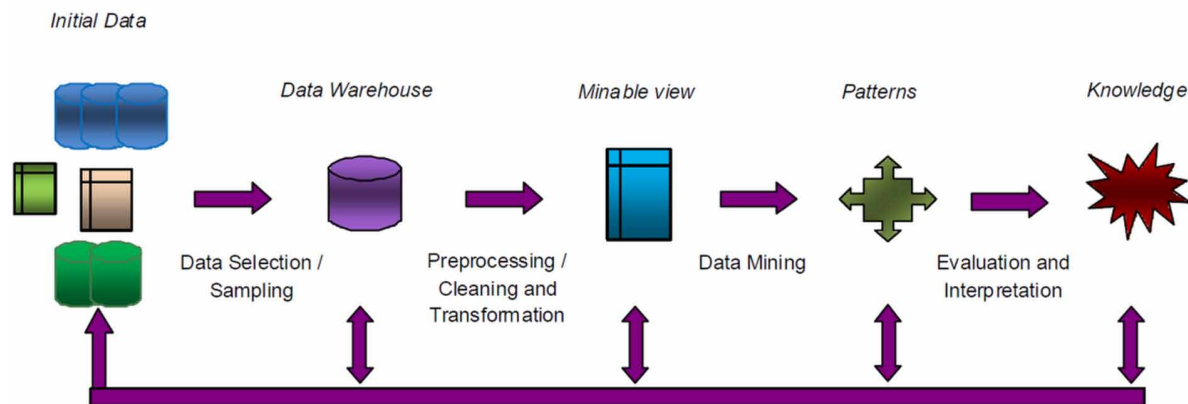
AI comprises not only Machine Learning but other disciplines such as Robotics, Logic Programming and Inductive Logic Programming (ILP), which is a field of Machine Learning and Data Mining. Figure 1 shows a schematic view of the relationships between Data Mining, Machine Learning, AI and other associated disciplines.

DATA MINING AND THE KDD

Several definitions of Data Mining can be found in the literature. Witten and Frank (2011) refers to Data Mining as the process of extraction of previously-unknown, useful and understandable knowledge from big volumes of data, which can be in different formats and come from different sources. Hernández-Orallo, Ferri and Ramírez-Quintana (2004) define Data Mining as the process of converting data into knowledge. However, there is still some confusion in the use of the terms ‘KDD’ and ‘Data Mining’. Although they are strongly related, they are not the same. Data Mining is a phase in the KDD process, an iterative process that encompasses other phases or activities as the selection, cleaning and transformation of data, coming not only from traditional databases but also from other heterogeneous sources, such as plain text, data warehouses, images, sound, etc. The application of data mining algorithms to these data is achieved to discover valid, novel, potentially useful, and understandable hidden patterns. Once patterns are obtained, they are interpreted and evaluated, going back to previous phases for a new iteration if necessary.

Figure 2 shows a graphical representation of the KDD process. Some authors decompose KDD into other phases; nevertheless, its activities are basically the same.

Figure 2. The KDD process



During the Data Selection/Sampling phase, activities that have to do with data collection and integration take place. Frequently, data are collected from the most diverse sources hence they do not share the same format and structure. In these cases, an activity of data integration is required. This has given place to the concept of data warehouse, a technology used by organizations to collect data from different sources for the purpose of analysis and decision making, e.g. evolution forecast, organizational analysis, and strategy design.

Unlike databases, mainly used by applications based in traditional processing of data, known as OLTP (On-Line Transaction Processing), data warehouses are operated by OLAP (On-Line Analytical Processing) tools. Data warehouses have a different data organization from databases, adopting a multidimensional data model.

Like OLAP, Data Mining techniques can make use of data from data warehouses, but while OLAP tools allow aggregating detailed information, getting reports and views in real time and verifying patterns and hypothesis suggested by the user in an essentially deductive process, Data Mining methods are essentially inductive process that uses data to infer patterns. Nevertheless, OLAP tools can also be used in the first stages of the KDD process for a better understanding of data.

A second phase of Preprocessing/Cleaning and Transformation is sometimes needed. During this stage, a minable view of the previously integrated data is created. A minable view is a subset of the integrated data that will be input to the data mining algorithms; it eliminates the unnecessary data for the data mining task. To that effect, activities such as the creation of new attributes, selection of relevant attributes, attribute transformation and reduction in the number of observations can also take place. Sometimes, missing values and outliers (anomalous values) are also eliminated, especially when the algorithm used is sensible to their presence.

Once the minable view is ready, the Data Mining phase can proceed. In this phase, decisions about the method or technique (e.g. decision trees) to adopt for solving a task (e.g. classification) must be taken. In this phase the algorithms to be applied to the minable view (e.g. CART, C4.5) must be chosen. In the next sections an overview of Data Mining's tasks, methods and algorithms is given.

The last phase, Evaluation and Interpretation of patterns, is where the model learnt from data mining is evaluated, after which a new iteration begins starting from one of the previous stages whenever it is necessary.

Evaluation is a very important process in KDD since it involves determining which of the models among several alternatives is going to be kept for solving the problem at hand. Therefore, techniques to evaluate how well the different methods work and to compare them are indispensable.

Model evaluation techniques and measures depend on the kind of task to solve (descriptive/predictive). For example, a typical evaluation measure for classification models is error rate/accuracy, defined as the proportion of incorrectly/correctly classified instances over the total number of instances.

The most simple evaluation technique for predictive tasks is simple validation, where the evidence is split in two sets, one for testing (test set) and the other for training (training set). This is done to guarantee that the evaluation measure obtained from the test set is an independent measure, avoiding too optimistic estimations. However, there exist most sophisticated techniques such as cross validation, n-fold cross validation, useful when the evidence is scarce, or bootstrapping (Efron and Tibshirani, 1993) (Kohavi, 1995).

In descriptive tasks, such as clustering and association rules, different measures are used. For example, coverage (or support) and accuracy (or confidence) are used for association rules.

TASKS AND METHODS OF DATA MINING

A task is normally defined as a work that is assigned to a person. In the realm of AI a task is work that will be assigned to a machine or computer, and a better definition would be a work that will be done by a computer program.

A method is a procedure, technique, or algorithm for doing something: in this case a task. The main method that AI and its different branches use to accomplish the tasks is Inductive Learning (IL). In the following sections both tasks and methods of Data Mining are detailed.

DATA MINING TASKS

In AI, as in many other disciplines, there exist families of problems. Each of these families, is called learning task and it can be solved by means of diverse methods and algorithms. Pattern Recognition like image recognition –including face recognition–, text recognition, etc. can be considered one of the main tasks. But Classification, regression, clustering, correlations and association rules, can also be recognized among the most important of these tasks.

As result of a task and according to the problem it solves, a prediction or a description can be obtained, giving place to predictive and descriptive tasks. Predictive tasks, as its name indicates, estimate unknown or future values, while descriptive tasks search hidden patterns in data as a way of explaining or describing them.

One of the most common predictive tasks is classification, while correlations, association rules, functional dependencies and clustering can be considered as descriptive tasks.

A classification task consists in labeling each new instance with a tag (the class), from a set of possible classes. Classification predicts the class of an instance from the values of its attributes, e.g. a classifier that has previously learned the concept of spam from old messages labeled as spam/no spam can be used to classify new messages.

There are variations to classification like categorization, where an instance can be labeled with more than one class. Other variants to classification are soft classification and probabilistic classification. A soft classifier not only assigns a class but also an estimation of the degree of certainty. Probabilistic classification is a generalization of soft classification, where for each instance the classifier predicts a probability distribution for the set of possible classes. When the number of classes are two, a probabilistic classifier is equivalent to a soft classifier.

Another descriptive task is preference learning, where a ranking is learnt. The learning process is carried out not from a set of instances like in classification but from a set of sequences of instances, where the order in a sequence represents the preference order of the instances in the sequence. For example, given a set of candidates for a job position, to get a preference order of the candidates based on their attribute's values and previous preference rankings from which to learn the new preference order.

Regression is also another important predictive task. In this case, a real value is predicted, i.e., a real function is learnt. An example of a regression task is the productivity estimation for a company based on past productivity values. When a real value is predicted for the future, this task is referred as estimation, and when the real value predicted is between two known values it is referred as interpolation.

On the other hand, correlation is a descriptive task. It is useful to determine whether two numeric variables are correlated, i.e. whether they have a similar behavior. When there is a positive correlation, both variables increase and decrease at the same time; if the correlation is negative one decreases when the other increases. A correlation not only can be positive or negative but it also can be strong or weak or nonexistent. This task is used in data mining to determine redundant attributes or dependencies among them.

Association rules is also a descriptive task that, as correlations, finds existing relationships among categorical (non numerical) attributes. Functional dependencies are a variation of association rules, where

only one attribute is determined from a subset of the remaining attributes. For example, a functional dependency could determine that a person is a homeowner from his/her age, income and marital status.

The descriptive task of clustering consists in partitioning a set of instances in natural groups called clusters. Originally, the central concept for clustering was the use of a numeric measure of similarity between the instances, creating clusters where the similarity among the members of a cluster is maximized and the similarity between the members of different clusters is minimized. A different approach to clustering is conceptual clustering proposed by Michalski (1980), which is not based in the notion of similarity but in conceptual cohesion (Michalski and Stepp, 1983). In conceptual clustering, two individuals belong to the same cluster not because they are similar (according to a similarity measure or distance) but because they are part of a same concept or, in other words, they are described by a same concept. These two different views of clustering have given place to different approaches to this task: similarity based clustering and conceptual clustering.

The main application of clustering is in guessing that new instances that have been considered members of a cluster have similar behavior or characteristics that the members of that cluster. For example, an on-line shop could make use of clustering to create groups of customers according to their purchase preferences, and then when a customer buys an article the system offers other articles based on the preferences of others customers in the same cluster.

DATA MINING METHODS AND ALGORITHMS

There is an important number of learning methods that support the previously described tasks. For example, a classification task can be solved using different methods, e.g. neural networks or decision trees. At the same time, each particular method can be implemented by different algorithms. For example, there are several decision trees algorithms, such as ID3 (Quinlan, 1986) or its extension C4.5 (Quinlan, 1993). Many of them can be used to solve different tasks, e.g. k-nearest neighbor algorithm can be used for classification, regression and also for clustering.

In general, DM methods can be categorized as supervised and unsupervised. Supervised methods learn models from examples, where there is a teacher or supervisor (expert) who defines the classes and provides examples for each class. These examples form a training set, which is used to build a model that allows predicting the class of new examples. Some supervised methods learn a function, i.e. the class are disjoint, while others learn a relationship, where classes can be overlapped. The first corresponds to the task of classification, while the second to categorization.

In unsupervised methods, learning is by observation and discovery; the examples used for learning are not labeled with a class as in supervised methods but the method must observe the examples and recognize patterns by themselves.

In general, both supervised and unsupervised methods adhere to one of two inductive learning approaches: symbolic learning and similarity-based learning. In the next section a description of both approaches and an overview of their main methods are given.

SIMILARITY-BASED LEARNING

Similarity-base learning methods share the concept of similarity and other associated concepts like the notion of distance and neighborhood or proximity. Some kind of mechanism is needed to compare two

objects and decide if they are similar. One way is by means of the use of a similarity function. A similarity function is a real function that given two objects returns a real value representing how similar these two objects are; it returns higher values while more similar the objects are. For example, a similarity function frequently used in Machine Learning is the cosine similarity measure between two vectors of real numbers; it is often used to compare documents.

The concept of distance or metric is associated to the notion of similarity, being the distance between two objects inversely proportional to the similarity between them. Even though a distance is a particular case of a similarity function, distances have certain properties that can be advantageously used by learning techniques.

Similarity-based methods support the idea that similar objects should behave in similar way. For example, given a similarity measure, a new instance could be classified as belonging to the class that has the most similar instance; or in a clustering task, it could be placed it into the cluster that has the instances more similar to it.

A brief description of some well-known similarity-based supervised and unsupervised methods follows.

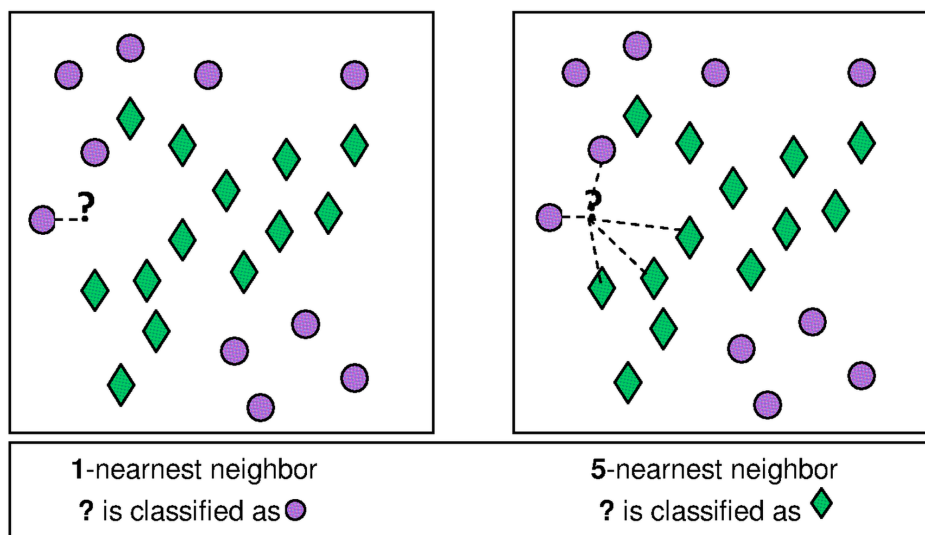
SIMILARITY-BASED SUPERVISED METHODS

Similarity-based supervised methods learn from a set of labeled instances employing a similarity function or a distance to make their predictions. Examples of them are k-Nearest Neighbors, Fisher's discriminant, Support Vector Machines, Learning Vector Quantization, among others.

k-NN (k-Nearest Neighbors)

It is very simple and probably the best-known supervised method. The idea behind k -NN is that the k nearest neighbors' majority class is assigned to each new instance. The main problem is determining the optimum value for k (see Figure 3). Usually, k is determined using heuristics; however, $k = \sqrt{n}$ where n is the number of examples, is an option with a theoretical base.

Figure 3. Difference between the nearest neighbor classifications in function of the chosen value of k



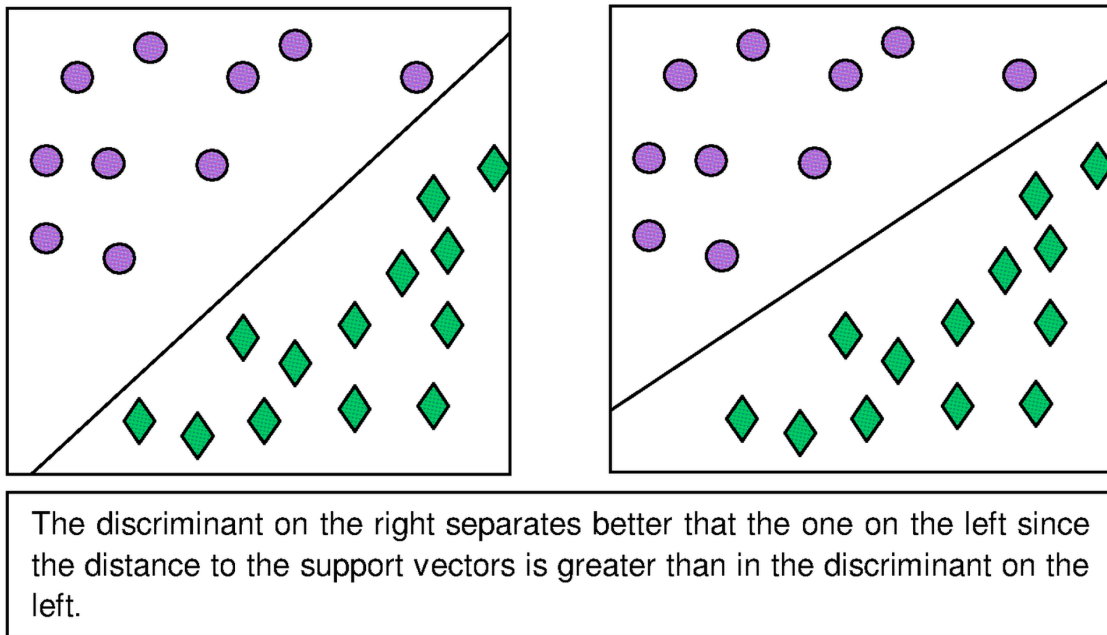
Any data representation can be used provided that a suitable similarity function is found.

Support Vector Machines (SVM)

This method searches a linear discriminant that maximizes the distance to the examples in the border of a class.

The classifier simply computes the hyperplane that perfectly separates the classes and that maximize the distance from the examples in the class border to the frontier (see Figure 4).

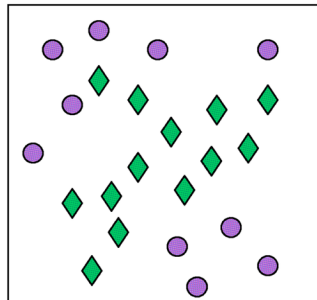
Figure 4. Linear discriminants



SVM are very efficient, even for hundred of dimensions because the linear discriminant has only to see a few points, called the support vectors, avoiding all the points that are not close to the frontier.

When data cannot be lineally discriminated (see Figure 5), a *kernel* function can be applied, augmenting the number of dimensions in such a way that the data are separable.

Figure 5. Set of examples not lineally separable



A SVM model is a representation of the examples as points in a space, transformed in such a way that the examples are clearly separated. When a new example must be classified, this is transformed by the kernel function into the same space and classified according to the side of the hyperplane where the new example falls.

Learning Vector Quantization (LVQ)

The learned model is a collection of prototypes, where a new example is classified according to its proximity to those prototypes. A prototype of a set of elements is the element of the set that is in the center of the elements according to its metric space; i.e. it minimizes the sum of the distances to the other elements in the set. When this element does not belong to the set, it is called centroid.

Given a training set with k classes, the algorithm randomly choose k prototypes, whose positions change until a given threshold is reached. This is an iterative process, where, firstly, an example from the training set is chosen and labeled with the class of the closer prototype. Next, the prototype of the class is recomputed. Depending on the labels of the prototype and the example, the prototype position will change, making closer the new prototype to the example if both labels match or the opposite if they do not match. The process stops when the change in the prototypes positions is lesser than a give threshold or there are no more training examples.

SIMILARITY-BASED UNSUPERVISED METHODS

In similarity-based unsupervised learning, the evidence is presented as a set of unlabeled elements and the objective is to find a descriptive model from the evidence, in this case using a similarity function.

Learning of descriptive models can be carried out in different ways and by multiple tasks, e.g. clustering, correlations and factorizations, association rules and functional dependencies.

A brief description of similarity-based clustering methods is given below; the presented clustering methods have been classified into hierarchical clustering and partitioning clustering, a common classification of clustering methods.

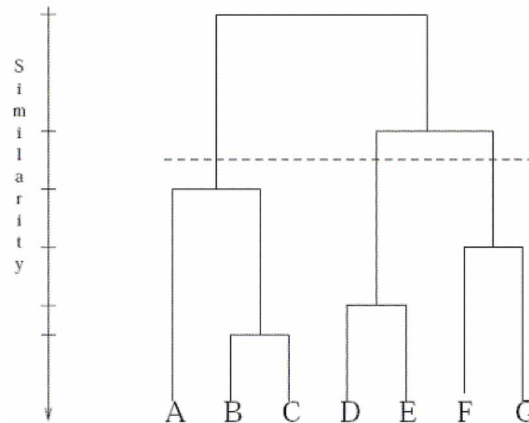
Hierarchical Clustering

These algorithms produce nested series of partitions of the evidence using as criterion, for splitting or merging the clusters, a similarity function. Well-known examples of hierarchical clustering are agglomerative and divisive hierarchical clustering.

A hierarchical algorithm produces as result a dendrogram, which represents the nested clusters. It also shows the similarity levels where a clustering changes to a new clustering (see Figure 6).

A dendrogram can be decomposed at different levels obtaining different data partitions. For example, in Figure 6, the dotted line cuts the hierarchy at the level corresponding to the clustering formed by the groups $\{A, B, C\}$, $\{D, E\}$ and $\{F, G\}$.

Figure 6. Example of dendrogram produced by a hierarchical clustering algorithm



Partitioning Clustering Algorithms

Unlike hierarchical clustering algorithms, which create a clustering hierarchy, these algorithms produce only one partition of the evidence.

One of the most well-known partitioning algorithms is k-means. Its process begins with an initial random partition formed by k clusters and proceeds reassigning the examples to other clusters driven by the distance between the example and the clusters' centroids. Once all the examples have been reassigned, the clusters' centroids are recomputed. This process continues until a convergence criterion is reached. Although k-means is popular because its implementation is simple, it has the drawback of being sensitive to the selection of the initial partition.

MODEL-BASED SYMBOLIC LEARNING

A weakness of similarity-based methods is the lack of interpretability of the discovered models, which are determined by the similarity between the evidence elements. This makes these models hard to understand by humans. However, in many cases to have an explicative model is a must; for example, a learned model should provide an explanation or description of its recommendations when refusing a loan to a customer, going beyond a label that classifies the customer as unreliable. It is here where symbolic methods are useful. In the next two sections some of these methods are described.

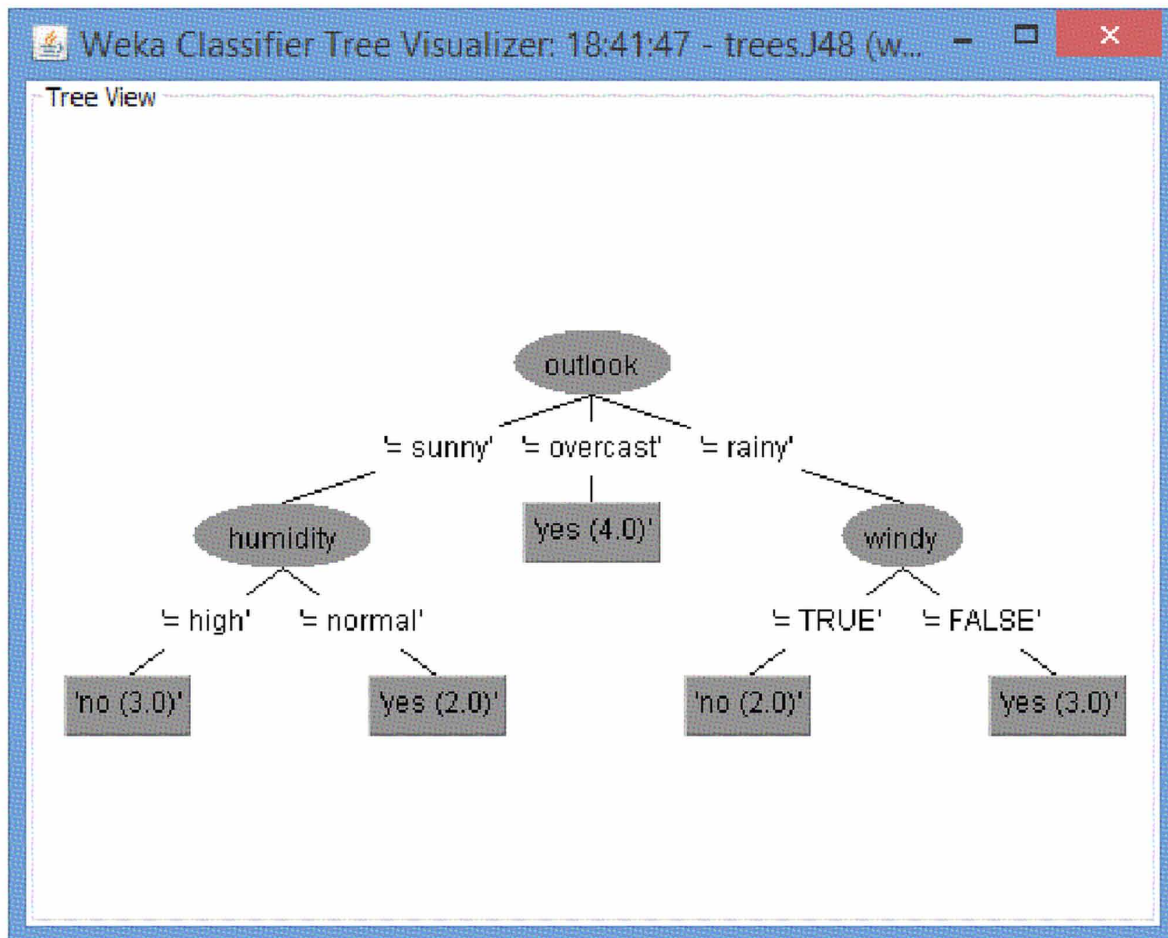
MODEL-BASED SYMBOLIC SUPERVISED METHODS

Possibly, the most well-known supervised symbolic method is decision trees, where a predictive model in the shape of a tree structure is learnt from the evidence. Decision tree branches represent attributes conjunctions and its leaves correspond—in the case of classification trees—to a set of labels or classes, or to real values in the case of regression trees.

Figure 7 shows a decision tree produced by the J48 algorithm in the Weka toolset (Weka, n.d.) induced from its weather dataset. Given a new instance, it allows ascertaining whether to play a tennis

match. The tree shows in its intermediate nodes the attributes outlook, humidity, and windy with their corresponding values on its arcs and the values yes/no of the class (play), on the leaves.

Figure 7. Decision tree to determine whether to play tennis



Whenever a new instance must be classified, the tree is traversed from the root according to the attribute values in the successive nodes until a leaf is reached and the instance is classified with the value in the leaf.

MODEL-BASED SYMBOLIC UNSUPERVISED METHODS

Two representative unsupervised symbolic methods are Conceptual Clustering (Michalski, 1980), (Michalski and Stepp, 1983) and Formal Concept Analysis (FCA) (Ganter et al., 1999).

Conceptual Clustering

Classical distance-based unsupervised learning methods do not provide conceptual descriptions of the discovered clusters. This is a limitation of these methods that leave to the user the problem of interpreting clusters.

Michalski introduced, in 1980, the idea of grouping objects in categories that are described by concepts. As it is explained in (Guerra-Sandón et al., 2012), the problem addressed by conceptual algorithms is of great practical usefulness. They are not limited to produce a list of objects that are in the same cluster and hence share similar characteristics or properties, but the intention of conceptual clustering is to provide additional information by giving the common properties of the elements in each cluster.

Michalski was the precursor of these ideas and nowadays there are several algorithms for conceptual clustering. A conceptual clustering classification is given in (Guerra-Sandón et al., 2012). There also exists hybrid proposals like HDCC (Funes et al., 2008), where an integration of both paradigms is achieved.

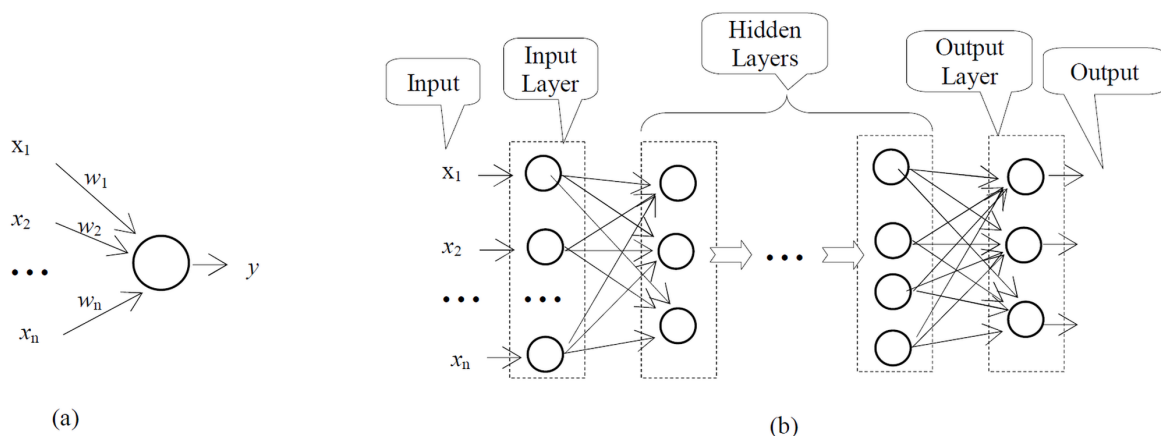
Formal Concept Analysis (FCA)

FCA allows identifying conceptual structures from data. These structures correspond to lattices of formal concepts, where each formal concept is a pair (set of properties, set of examples) and where the set of properties in the formal concept describes completely its set of examples. These lattices of formal concepts consist of a partial order relation defined as a generalization relation between formal concepts. The main objective of FCA is offering compact descriptions of groups of data rather than a data clustering.

NEURAL NETWORKS

Artificial Neural Networks (ANNs) have a long history, quite as long as AI itself. They were inspired by the functioning of biological neurons. Basically, neurons in a brain are interconnected by axons. A neuron receives inputs from other neurons and when those inputs exceed a threshold the neuron sends an output to other neurons. Artificial Neurons (AN) mimic the functions of 'real' neurons. Each AN is

Figure 8 (a) Perceptron (single layer). (b) Multiple Layer ANN



a mathematical function $y = f(\sum_i (x_i w_i))$ where x_i are the inputs and w_i the associated weights. Figure 8 (a) shows a particular ANN, called Perceptron, which has one AN its inputs, weights and its output.

The origins of ANNs can be found in the work of McCulloch & Pitts, (1943), who first constructed a theory of AN that gave way to the Threshold Logic Unit (TLU), or Linear Threshold Unit. Later, Rosenblatt (1957) created the Perceptron. Initially, it was intended as an algorithm for supervised learning of binary classifiers, later it was implemented as a machine for perceiving and recognizing images. The Perceptron was a single layer ANN. Later Minsky & Papert (1969) showed that single layer ANNs –like the Perceptron– are unable to perform a logical xor.

The next advance was the creation of ANNs with multiple layers, i.e. ANNs that have an input layer, an output layer and one or more intermediate layers called hidden layers since they cannot be ‘seen’ from the outside (see Figure 8 (b)).

An ANN can be trained to accomplish a task. An important improvement in ANN training was the introduction of the technique of backpropagation, an efficient method for calculating the weights updates to improve the network until it is able to perform the task for which it is being trained (Goodfellow et al., 2016). There is a bit of a controversy on the discovery of the method but some of the pioneers in the area are Kelley (1960), Bryson (1961), and then Mizutani, Dreyfus, & Nishio (2000).

Learning algorithms for ANNs can be supervised or unsupervised. When the desired output is given to the ANN input layer is supervised. Unsupervised learning provides unlabeled data and the algorithm tries to make sense of it by extracting features and patterns on its own. There are also semi-supervised learning algorithms, where normally only a small labelled set of data is provided against a large set of unlabelled data.

A more recent and growing branch of Machine Learning is Deep Learning, which implies the use of ANNs with multiple layers. These neural architectures come in a number of varieties such as convolutional neural networks, deep neural networks, deep belief networks, recurrent neural networks and others. They are applied to many fields e.g. image and audio recognition, natural language processing –including machine translation– social network research, bioinformatics, drug design, image analysis, game programs, etc., where they produced results comparable to and in some cases superior to human experts. ANNs for Deep Learning are sometimes called Deep Neural Networks (DNN). DNNs are huge ANNs, whose current upper number is in the range of 100-160 billion of ANs.

ISSUES, SOLUTIONS, AND RECOMMENDATIONS

Along the last the years several issues regarding Ethics, mainly related to self-driving vehicles and autonomous weapons, both areas where AI is being used, have arisen.

As Maxmen (2018) said “When a driver slams on the brakes to avoid hitting a pedestrian crossing the road illegally, she is making a moral decision that shifts risk from the pedestrian to the people in the car”. So that moral burden on a human driver shifts now to the software and hardware in a self driving vehicle. Considering that the software and hardware are built by humans, some of the questions are: what kind of moral decision is built, programmed, into the AI program that drives the vehicle, since there are situations, around a particular maneuver of the vehicle, that may involve a moral decision? Who is legally responsible for the damages –including loss of life– to third parties involved (passengers and bystanders included) in an accident of a self-driving vehicle? These questions will necessary have to be answered.

Also the building and deploying of autonomous weapons is of great concern and has lead to a debate on the subject by prominent researches in the AI field, exposing pros and cons clearly. For more infor-

mation on this, see Scharre & Horowitz, (2015), Russell, Tegmark & Walsh, (2015), Ackerman, (2015), Gubrud, (2016), Guizzo & Ackerman, (2016), Scharre, (2017), Russell, Aguirre, Conn & Tegmark, (2018).

Another important problem intimately associated with Data Mining and AI is privacy and protection of personal data. While these disciplines have benefited from the existence of large volumes of data from diverse sources, these data are not always legally obtained. The laws and regulations related to privacy and data protection varies around the world and are constantly changing in the search of a balance between rights of individuals and other interests.

FUTURE RESEARCH DIRECTIONS

More and more are the areas where AI and Data Mining are having great impact and where some related research is being conducted. Below a short sample of what can be considered part of the future research directions on the field of AI is given.

Companies such as Jobscan (2019) that provides job seekers with help in improving their resume, claim –based on their own data– that over 98% of Fortune 500 companies use Applicant Tracking Systems (ATS), based on AI. Also Korn Ferry (2018), a consulting firm in Los Angeles, USA, states something similar.

Using the concept of Neuromorphic, proposed by Mead (1990), Intel is planning to build a Neuro-morphic System of 8 Million Neurons and to reach 100 Million by 2020, Moore, (2019).

Javed et al (2019) use AI and Robotics to help children with Autism Spectrum Disorder.

Not only image and face recognition is a growing area of research, but also recognition of fake images is also an area where AI is being used, see Leopold, (2019).

CONCLUSION

Although AI has become a mature discipline and Data Mining has contributed greatly to technological development in recent years, the question remains: Is building a human-like computer a valid goal of AI? This question intends to be answered by proposals such as technological singularity, sometimes referred only as singularity, a concept first put forward by Vinge (1993). Basically the concept holds that the technological progress will eventually lead to the creation of a superhuman intelligence and the consequently demise of the human era. Or Technium is another term, in this case coined by Kelly (2010), used “to designate the greater, global, massively interconnected system of technology vibrating around us” Kelly (2010, p. 11). This notion has been criticized as having a certain religious bias, but also by the continuous application of AI to different areas that were before reserved for humans.

Beyond the aforementioned philosophical question, AI continues to occupy places previously reserved only for man. This leads us to rephrase the question: Is it in this trend that the answer can be found?

REFERENCES

Ackerman, E. (2015). We Should Not Ban ‘Killer Robots’, and Here’s Why. *IEEE Spectrum*, 29. <https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/we-should-not-ban-killer-robots>

Bryson, A. E. (1961). A gradient method for optimizing multi-stage allocation processes. In *Proceedings of the Harvard University, Symposium on digital computers and their applications*. Harvard University Press.

Cohen, Paul R. (2006). If Not Turing's Test, Then What? *AI Magazine*, 26(4).

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall. doi:10.1007/978-1-4899-4541-9

Ernst, G. W., & Newell, A. (1969). *GPS: A Case Study in Generality and Problem Solving*. Academic Press.

Funes, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2008). *Hierarchical Distance-based Conceptual Clustering. Machine Learning and Knowledge Discovery in Databases, LNAI 5211/2008*. Springer.

Ganter, B., Stumme, G., & Wille, R. (1999). *Formal Concept Analysis*. Springer. doi:10.1007/978-3-642-59830-2

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Gubrud, M. (2016). Why Should We Ban Autonomous Weapons? To Survive. *IEEE Spectrum*. <https://spectrum.ieee.org/autaton/robotics/military-robots/why-should-we-ban-autonomous-weapons-to-survive>

Guerra-Sandón, A., Vega-Pons, S., & Ruiz-Shulcloper, J. (2012). Algoritmos de Agrupamientos Conceptuales: un estado del arte. *Reporte Técnico CENATAV*.

Guizzo, E., & Ackerman, E. (2016, June). When Robots Decide to Kill. *IEEE Spectrum*.

Hernández-Orallo, J. (2000, October). Beyond the Turing Test. *Journal of Logic Language and Information*, 9(4), 447–466. doi:10.1023/A:1008367325700

Hernández-Orallo, J. (2019). Unbridled mental power. *Nature Physics*, 15, 106. www.nature.com/naturephysics

Hernández-Orallo, J., Ramírez-Quintana, M. J., & Ferri, C. (2004). *Introducción a la Minería de Datos*. Pearson Education S.A.

Javed, H., Burns, R., Jeon, M., & Howard, A. M. (2019). *A Robotic Framework to Facilitate Sensory Experiences for Children with Autism Spectrum Disorder: A Preliminary Study*. <https://arxiv.org/ftp/arxiv/papers/1901/1901.00885.pdf>

Jobscan. (2019). <https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/>

Kelley, H. (1960, October). Gradient Theory of Optimal Flight Paths. *ARS Journal*, 30(10), 947–954. doi:10.2514/8.5282

Kelly, K. (2010). *What Technology Wants*. Penguin Group.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann.

Korn Ferry. (2018). *Korn Ferry Global Survey: Artificial Intelligence (AI)*. Reshaping the Role of the Recruiter. <https://www.kornferry.com/press/korn-ferry-global-survey-artificial-intelligence-reshaping-the-role-of-the-recruiter>

Leopold, G. (2019). Facing Up to Image Fakery. *Datanami*. <https://www.datanami.com/2019/06/21/facing-up-to-image-fakery/>

Maxmen, A. (2018). Self-driving car dilemmas reveal that moral choices are not universal. *Nature*, 562(7728), 469–470. doi:10.1038/d41586-018-07135-0 PMID:30356197

Mcculloch, W. S., & Pitt, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity*. *Bulletin of Mathematical Biology* Vol. 52, No. 1/2. pp. 99-115. 1990. Pergamon Press plc. Society for Mathematical Biology. *. *Reprinted from the Bulletin of Mathematical Biophysics*, 5, 115–133. doi:10.1007/BF02478259

Mead, C. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE*, 78(10), 1629-1636. <http://resolver.caltech.edu/CaltechAUTHORS:20141222-113217405>

Michalski, R. S. (1980). Knowledge acquisition through conceptual clustering. *Policy Analysis and Information Systems*, 4, 219–244.

Michalski, R. S., & Stepp, R. E. (1983). Learning from Observation: Conceptual Clustering. In J. G. Carbonell, R. S. Michalski, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 331–363). TIOGA Publishing Co.

Minsky, M., & Papert, S. (1969). *Perceptrons. An Introduction to Computational Geometry. Expanded Edition*. Massachusetts Institute of Technology.

Mizutani, E., Dreyfus, S., & Nishio, K. (2000). On derivation of MLP backpropagation from the Kelley-Bryson optimal-control gradient formula and its application. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 27-27 July 2000, Como, Italy*. IEEE.

Moore, S. K. (2019, July). Intel's Neuromorphic System Hits 8 Million Neurons, 100 Million Coming by 2020. *IEEE Spectrum*, 15. Retrieved July 2019, from <https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/intels-neuromorphic-system-hits-8-million-neurons-100-million-coming-by-2020>

Piatetsky-Shapiro, G., & Frawley, W. J. (Eds.). (1991). *Knowledge discovery in databases*. AAAI Press/MIT Press.

Quinlan, J. R. (1986, March). Induction of Decision Trees. *Machine Learning*, 1(1), 81–106. doi:10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Rosenblatt, F. (1957). *The Perceptron--a perceiving and recognizing automaton*. Report 85-460-1, Cornell Aeronautical Laboratory.

Russell, S., Aguirre, A., Conn, A., & Tegmark, M. (2018). Why You Should Fear “Slaughterbots” –A Response. *IEEE Spectrum*. <https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-you-should-fear-slaughterbots-a-response>

Russell, S., Tegmark, M., & Walsh, T. (2015). Why We Really Should Ban Autonomous Weapons: A Response. *IEEE Spectrum*. <https://spectrum.ieee.org/autaton/robotics/artificial-intelligence/why-we-really-should-ban-autonomous-weapons>

Russell, S. J., & Norvig, P. (2010). Artificial Intelligence - A Modern Approach (3rd ed.). Pearson Education, Inc.

Scharre, P. (2017). Why You Shouldn't Fear "Slaughterbots". *IEEE Spectrum*. <https://spectrum.ieee.org/autaton/robotics/military-robots/why-you-shouldnt-fear-slaughterbots>

Scharre, P., & Horowitz, M. C. (2015). Ban or No Ban, Hard Questions Remain on Autonomous Weapons. *IEEE Spectrum*. <https://spectrum.ieee.org/autaton/robotics/military-robots/ban-or-no-ban-hard-questions-remain-on-autonomous-weapons>

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 49(236), 433–460. doi:10.1093/mind/LIX.236.433

Vinge, V. (1993). *The Coming Technological Singularity: How to Survive in the Post-Human Era*. <https://www.semanticscholar.org/paper/The-coming-technological-singularity%3A-How-to-in-the-Vinge/3ab2d953596fea9131e622410c64b3d114a84f0c>

Weizenbaum, J. (1976). *Computer Power and Human Reason*. W. H. Freeman and Company.

Weka. (n.d.). Retrieved from the Weka Site: <https://www.cs.waikato.ac.nz/ml/weka/>

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining. Practical Machine Learning, Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

ADDITIONAL READING

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27–34. doi:10.1145/240455.240464

Hernández-Orallo, J. (2017). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press., doi:10.1017/9781316594179

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning, with Applications in R*. Springer-Verlag New York., doi:10.1007/978-1-4614-7138-7

Kurzweil, R. (2005). *The Singularity is Near*. Viking Books.

Tegmark, M. (2017). *Life 3.0: being human in the age of artificial intelligence*. Alfred A. Knopf.

KEY TERMS AND DEFINITIONS

AI4Good: By building lasting communities that bring the best technologies to bear on the world's most important challenges, the AI for Good Foundation drives forward solutions that support the United Nations Sustainable Development Goals (SDG's). We do this by coordinating the AI research commu-

nity, technologists, data, and infrastructure with the stakeholders on the ground, policy makers, and the broader public (<https://ai4good.org/about/>).

Automated Reasoning: To use stored information to answer questions and to draw conclusions.

Classification: Inductive task where a predictive model is learnt from objects labeled with a class and whereby it is possible to predict the class of new objects.

Cloud Computing: Generally refers to describe data centers available to many users over the Internet that provide data storage and computing power. It's sometimes used to process DNNs.

Clustering: Inductive task where a set of unlabeled objects is partitioned into groups (clusters) and where objects in a same cluster have similar characteristics, maximizing the similarity intra cluster and minimizing the similarity inter cluster.

Inductive Learning: Induction is the inference of information from data and inductive learning is a model building process where the data are analyzed to find hidden patterns.

Machine Learning: The study and use of algorithms and mathematical and statistical models in computer systems to perform a specific tasks but not using explicit instructions, on the contrary relying on discovery of patterns and inference. Is considered part of Artificial Intelligence. The software is to adapt to new circumstances and to detect and extrapolate patterns.

Natural Language Processing: The software is to be able to communicate successfully in a Natural Language (e.g., English).

Robotics: The machine should be able to manipulate objects and move about.

Weak AI: Also known as Narrow AI is AI aimed at a particular task.

Big Data Mining Algorithms

M. Govindarajan

Annamalai University, India

INTRODUCTION

Data mining is widely used in fields such as science, engineering, medicine, and business. With this technique, previously hidden insights have been unearthed from large amounts of data to benefit the business community (D. Che et al., 2013). Since the establishment of organizations in the modern era, data mining has been applied in data recording. For flexible data analysis, Begoli and Horey (2012) proposed three principles: first, architecture should support many analysis methods, such as statistical analysis, machine learning, data mining, and visual analysis. Second, different storage mechanisms should be used because all of the data cannot fit in a single type of storage area. Additionally, the data should be processed differently at various stages. Third, data should be accessed efficiently. To analyze Big Data, data mining algorithms that are computer intensive are utilized. Such algorithms demand high-performance processors. Many algorithms were defined earlier in the analysis of large data set. The purpose of this chapter is to provide an analysis of different machine learning algorithms available for performing big data analytics.

BACKGROUND

C. Yadav et al., (2013) present a review of some old algorithms that can handle large data set as Nearest Neighbor Search, Decision Tree and Neural Network. A. Fahad et al., (2014) present a survey of existing clustering algorithms of different categories (Partitioning-based, Hierarchical-based, Density-based, grid-based and model based). Tsai et al., (2015) begins with a brief introduction to data analytics followed by the discussions of big data analytics. Some important open issues and further research directions will also be presented for the next step of big data analytics. D.Rajkumar and S.Usha (2016) review on various big data mining algorithms and the methods employed to handle such a vast data is also discussed. J. Sangeetha and V. Sinthu Janita Prakash (2017) surveys about the big data mining techniques, data slicing techniques and clustering techniques and also discusses about its advantages and drawbacks. The performance and quality measurement of the big data mining algorithms, mining platforms, data slicing techniques and clustering techniques are discussed. Ripon Patgiri (2018) presents the taxonomy of Big Data and also present in-depth insight on the Big Data paradigm. Mounica Vennapusa et al., (2019) give the survey on applications and challenges of Machine Learning techniques, advanced learning methods towards Bigdata.

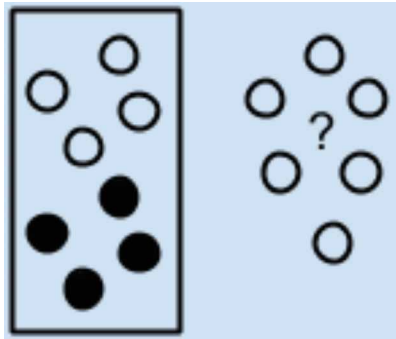
FOCUS OF THE ARTICLE

Machine learning is a field of computer science that try to enable computers with the ability to learn without explicitly programming them. The purpose of this chapter is to provide an analysis of different Machine learning algorithms available for performing big data analytics. The ML algorithms are categorized in three key categories, namely, supervised, unsupervised and semi-supervised ML algorithm. The Big Data Analytics or Big Data Security analytics fully depends on Data Mining, where the Machine Learning techniques are subset of Data Mining (Ripon Patgiri, 2018).

Supervised Learning Algorithms

The supervised learning algorithms are trained with a complete set of data and thus, the supervised learning algorithms are used to predict/forecast.

Figure 1. Supervised Learning Algorithms

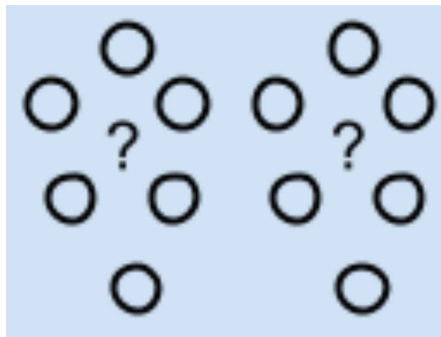


Example problems are classification and regression. Example algorithms include: Logistic Regression and the Back Propagation Neural Network.

Unsupervised Learning Algorithms

The unsupervised learning algorithms starts learning from scratch, and therefore, the unsupervised learning algorithms are used for clustering.

Figure 2. Unsupervised Learning Algorithms

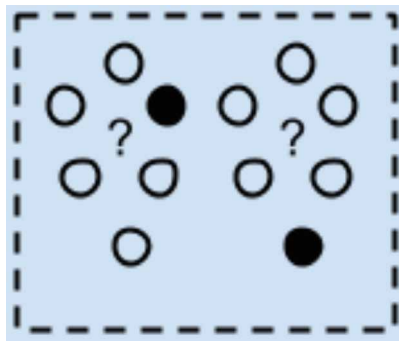


Example problems are clustering, dimensionality reduction and association rule learning. Example algorithms include: the Apriori algorithm and K-Means.

Semi-Supervised Learning Algorithms

The semi-supervised learning combines both supervised and unsupervised learning algorithms. The semi-supervised algorithms are trained, and the algorithms also include non-trained learning. Example problems are classification and regression.

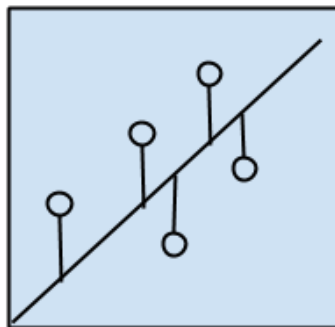
Figure 3. Semi-supervised Learning Algorithms



The following brief list identifies the different Machine Learning algorithms when applied in Big Data.

Regression algorithm: It predict output values using input features of the data provided to the system.

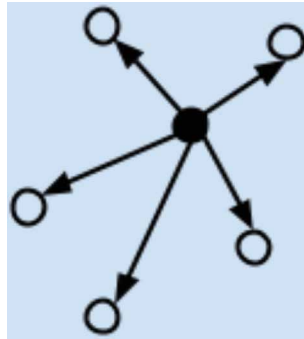
Figure 4. Regression Algorithms



Most popular algorithms under it are Linear Regression, Logistic Regression Multivariate Adaptive Regression Splines (MARS) and Locally Estimated Scatter plot Smoothing (LOESS).

Instance based algorithm: It compares new problem instances with instances in training, that are stored in memory.

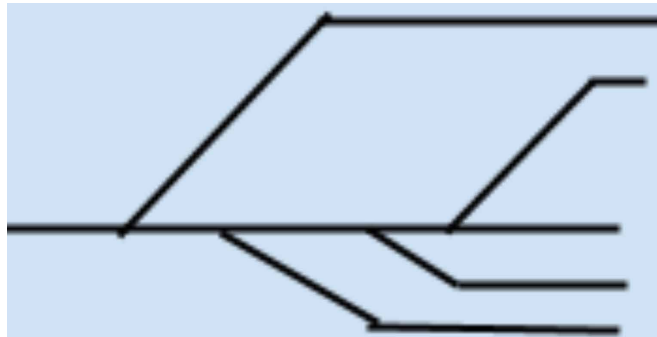
Figure 5. Instance-based Algorithms



Common algorithms can be k-Nearest Neighbor (kNN), Learning Vector Quantization (LVQ), Self-Organizing Map (SOM) and Locally Weighted Learning (LWL).

Regularization algorithm: It is a process of providing additional information to prevent overfitting or solve an ill-posed problem.

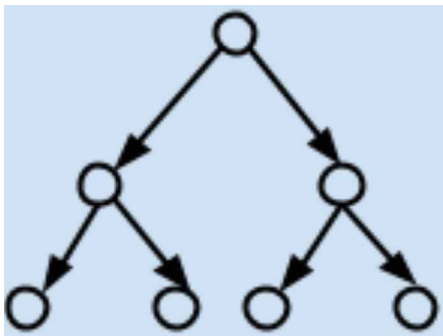
Figure 6. Regularization Algorithms



Algorithms which are common under it are Least Absolute Shrinkage and Selection Operator (LASSO), Least-Angle Regression (LARS), Elastic Net and Ridge Regression.

Decision tree algorithm: It solves the problem using tree representation.

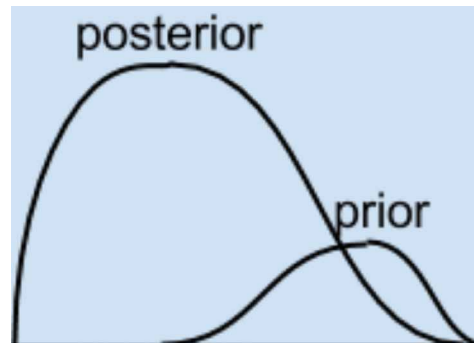
Figure 7. Decision Tree Algorithms



Some popular algorithms are Classification and Regression Tree (CART), C4.5 and C5.0, M5, and Conditional Decision Trees.

Bayesian algorithm: It is based on Bayesian method.

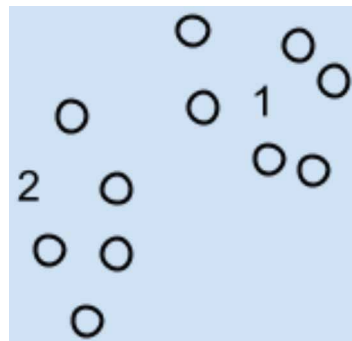
Figure 8. Bayesian Algorithms



Popular algorithms are Naive Bayes, Gaussian Naïve Bayes, Bayesian Network (BN), and Bayesian Belief Network (BBN).

Clustering algorithm: Grouping of data points based on similar features.

Figure 9. Clustering Algorithms



Some popular algorithms are KMeans, K-Medians, and Hierarchical Clustering.

Association Rule Learning Algorithms: It is used to discover relationship between data points.

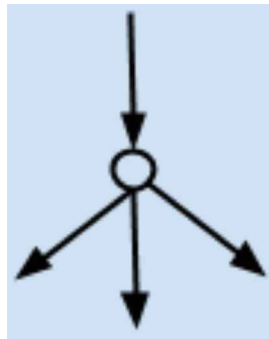
Figure 10. Association Rule Learning Algorithms



Some common algorithms under it are Apriori algorithm, and Eclat algorithm.

Artificial Neural Network Algorithms: It is based on working of biological neural networks.

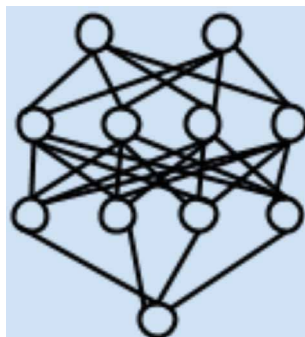
Figure 11. Artificial Neural Network Algorithms



Example of some popular algorithms are Back-Propagation Neural Network (BPNN), Perceptron, and Radial Basis Function Network (RBFN).

Deep Learning Algorithms: It uses unsupervised learning to set each level of hierarchy of features using features discovered at previous level.

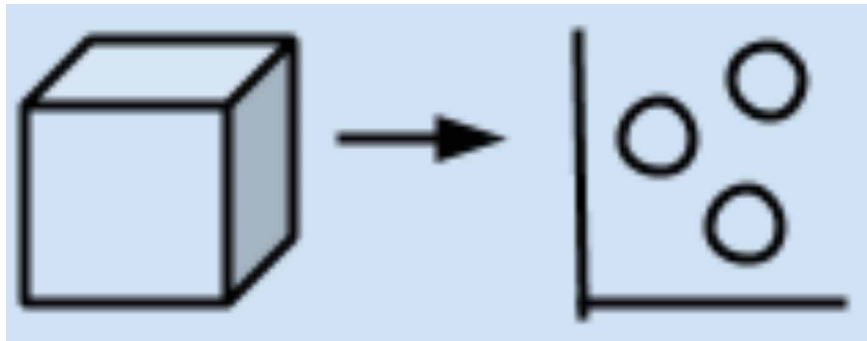
Figure 12. Deep Learning Algorithms



It has some popular algorithms such as Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), and Convolution Neural Network (CNN).

Dimensionality Reduction Algorithms: They reduce the number of features by obtaining a set of principal variables.

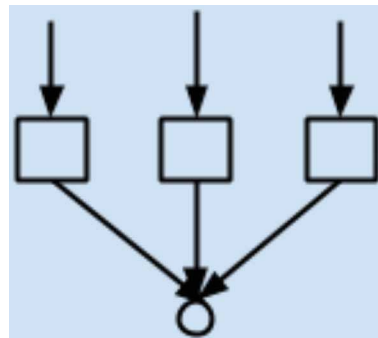
Figure 13. Dimensionality Reduction Algorithms



Some algorithms commonly under it are Principal Component Analysis (PCA) and Principal Component Regression (PCR), Partial Least Squares Regression (PLSR) and Linear Discriminant Analysis (LDA), Mixture Discriminant Analysis (MDA) and Quadratic Discriminant Analysis (QDA) and Flexible Discriminant Analysis (FDA).

Ensemble Algorithms: It combines multiple learning algorithms to obtain better predictive performance.

Figure 14. Ensemble Algorithms



Some well known algorithms are Random Forest, Gradient Boosted Regression Trees (GBRT), and Gradient Boosting Machines (GBM).

SOLUTIONS AND RECOMMENDATIONS

Some articles authored by IT specialists suggested practical computer science solutions, namely the development of discrimination-aware methods to be applied during the development of the algorithmic models. These techniques include: pre-processing methods that involve the sanitization or distortion of the training data set to remove possible bias in order to prevent the new model from learning discriminatory behaviours (e.g. (Hajian S et al., 2014 and Kamiran F et al., 2013); in-processing techniques that provide for the modification of the learning algorithm through the application of regularization to probabilistic discriminative models (Kamiran F et al., 2013) such as the inclusion of sensitive attributes to avoid discriminatory predictions (Pope DG et al., 2011 and Zliobaite I et al., 2016) or the addition of randomness to avoid overfitting or hidden model bias (Kroll JA et al., 2017); post-processing methods that involve the auditing of the extracted data mining models for discriminative patterns and eventually their sanitization (Hajian S et al., 2015). Practical suggestions to reinforce transparency in data mining include the development of interpretable algorithms that will give explanations on the logical steps behind a certain classification (Kroll JA et al., 2017 and Schermer BW, 2011), and the creation of transparent models that will allow individuals to see in advance how their behavior and choices will be interpreted by the algorithm or the infrastructure (Citron DK et al., 2014 and Hildebrandt M et al., 2010).

FUTURE RESEARCH DIRECTIONS

This Chapter reviews about the various big data mining algorithms. To support big data mining, high-performance computing platforms are required. It is understood that interestingness of discovered patterns, developing a global unifying theory, building efficient mining platforms or algorithms, privacy preserving, security, trust and data integrity are the major challenging issues in the current big data mining scenario. It is known that big data mining is an emerging trend in all science and engineering domains and also a promising research area. In spite of the limited work done on big data mining so far, it is believed that much work is required to overcome its challenges related to the above mentioned issues. (Sherin A et al., 2014)

CONCLUSION

Big data is gaining huge significance in the current scenario and consequently, big data mining emerged as an innovative and potential research area. Many algorithms were defined earlier in the analysis of large data set. Organizations often face teething troubles with respect to creating, managing, and manipulating the rapid influx of information in large datasets. Given the increase in data volume, data sources have increased in terms of size and variety. Data are also generated in different formats (unstructured and/or semi structured), which adversely affect data analysis, management, and storage. This variation in data is accompanied by complexity and the development of additional means of data acquisition. The extraction of valuable data from large influx of information is a critical issue in Big Data. Qualifying and validating all of the items in Big Data are impractical; hence, new algorithms must be developed.

REFERENCES

- Begoli, E., & Horey, J. (2012). Design principles for effective knowledge discovery from big data. *Proceedings of the 10th Working IEEE/IFIP Conference on Software Architecture (ECSA '12)*, 215–218.
- Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data Mining: challenges, issues, and opportunities. In *Database Systems for Advanced Applications* (pp. 1–15). Springer. doi:10.1007/978-3-642-40270-8_1
- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review* (Seattle, Wash.), 89, 1.
- Fahad, A., Alshatri, N., Tari, Z. A. L., Amri, A., Zomaya, A. Y., Khalil, I., Sebt, F., & Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3), 267–279. doi:10.1109/TETC.2014.2330519
- Hajian, S., Domingo-Ferrer, J., & Farras, O. (2014). Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 28(5–6), 1158–1188. doi:10.1007/10618-014-0346-1
- Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2015). Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery*, 29(6), 1733–1782. doi:10.1007/10618-014-0393-7
- Hildebrandt, M., & Koops, B.-J. (2010). The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review*, 73(3), 428–460. doi:10.1111/j.1468-2230.2010.00806.x
- Kamiran, F., Zliobaite, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3), 613–644. doi:10.1007/10115-012-0584-8
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. L. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.
- Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal. Economic Policy*, 3(3), 206–231. doi:10.1257/pol.3.3.206
- Rajkumar, D., & Usha, S. (2016). A Survey on Big Data Mining Platforms, Algorithms and Handling Techniques, *International Journal For Research In Emerging Science and Technology*, 3(1). NCRTCT, 16, 50–55.
- Ripon Patgiri. (2018). *Taxonomy of Big Data: A Survey*. arxiv:1808.08474v2 [cs.DC]
- Sangeetha, J., & Sinthu Janita Prakash, V. (2017). A Survey on Big Data Mining Techniques. *International Journal of Computer Science and Information Security*, 15(1), 482–490.
- Schermer, B. W. (2017). The limits of privacy in automated profiling and data mining. *Computer Law & Security Review*, 27(1), 45–52. doi:10.1016/j.clsr.2010.11.009
- Sherin, A., Uma, S., Saranya, K., & Saranya Vani, M. (2014). Survey on big data mining Platforms, algorithms and Challenges. *International Journal of Computer Science and Engineering Technology*, 5(9), 854–862.

Tsai, Lai, HanChao, & Vasilakos. (2015). Big data analytics: a survey. *Journal of Big Data*, 2(21).

Vennapusa & Bhyrapuneni. (2019). A Comprehensive Study of Machine Learning Mechanisms On Big data. *International Journal of Recent Technology and Engineering (IJRTE)*, 7(6S2), 773-779.

Yadav, C. Wang, S., & Kumar, M. (2013). Algorithm and approaches to handle large Data-A Survey. *International Journal of Computer Science and Network*, 2(3).

Zliobaite, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2), 183–201. doi:10.1007/10506-016-9182-5

KEY TERMS AND DEFINITIONS

Association Rule Learning Algorithms: It is used to discover relationship between data points.

Clustering Algorithm: Grouping of data points based on similar features.

Data Mining: Data mining is widely used in fields such as science, engineering, medicine, and business.

Decision Tree Algorithm: It solves the problem using tree representation.

Deep Learning Algorithms: It uses unsupervised learning to set each level of hierarchy of features using features discovered at previous level.

Ensemble Algorithms: It combines multiple learning algorithms to obtain better predictive performance.

Regression Algorithm: It predict output values using input features of the data provided to the system.

Semi-Supervised Learning Algorithms: The semi-supervised learning combines both supervised and unsupervised learning algorithms.

Supervised Learning Algorithms: The supervised learning algorithms are trained with a complete set of data and thus, the supervised learning algorithms are used to predict/forecast.

Unsupervised Learning Algorithms: The unsupervised learning algorithms starts learning from scratch, and therefore, the unsupervised learning algorithms are used for clustering.

Graph Database Management Systems: The Past, the Present, and the Future

Kornelije Rabuzin

Faculty of Organization and Informatics, University of Zagreb, Croatia

Martina Šestak

Faculty of Organization and Informatics, University of Zagreb, Croatia

INTRODUCTION

Nowadays, the increased amount and complexity of connected data stimulated by the appearance of social networks has shed a new light on the importance of managing such data, especially handling information about the connections. The most natural way of representing connected data is to represent them as nodes connected with relationships forming a graph. The concepts of graph theory have been used in many occasions to handle connected data in databases over the years.

The idea of storing data as a set of nodes and edges comprising a graph was implemented in various forms in data models used in the past. The network data model, developed in late 1960s, can be considered as the first data model, which most accurately incorporated this idea. However, it was not long before the relational data model appeared, and took over the entire database market for years, which it dominates even nowadays.

Even though several solutions have been introduced, which can be used to store graph-like data in available relational database management systems (RDBMSs), such as Oracle Graph or MariaDB OQ-Graph, NoSQL graph databases are considered to be the most efficient solution for storing such data, since the idea of storing highly connected data is their primary goal and purpose.

The possibilities of using graph databases in various application domains continue to grow. For instance, the most popular graph databases are Facebook Social Graph developed by Facebook to view connections between friends, or Amazon's graph-based recommendation system (Amazon Neptune service). Sieger discusses the possibilities of introducing concepts of graph databases to modern businesses, and how graph databases could be used for Supply Chain Management or finding deeper connections between patients with similar diseases (Sieger, 2016). Moreover, according to D. Woods, due to their ability to efficiently gain insight into connections between pieces of information, graph databases play (and will play) an important role in transforming modern businesses to data-driven organizations, which make use of their data through the concept of Master Data Management (Woods, 2015).

Since their beginnings, there have been many graph database management systems (GDBMSs) available on the database market (e.g., Neo4j, TitanDB, AllegroGraph, FlockDB, InfiniteGraph, etc.). Recently, new emerging trends can be observed on the GDBMS market; aside from Neo4j, which is still the most popular and constantly developed GDBMS, other „native“ GDBMSs are starting to be replaced and „outgrown“ by multi-model databases, such as OrientDB, ArangoDB, etc.

The aforementioned trends and solutions indicate that it is worth exploring how graph-like data has been stored and manipulated over the years in various data storage solutions, and which current data storage options are available to store and manipulate such data.

DOI: 10.4018/978-1-7998-3479-3.ch053

Therefore, the objective of this article is to give a timeline overview of developed graph data storage solutions in order to gain insight into past, present and future trends of GDBMSs. Additionally, throughout this article, the most influential factors and reasons for changes in trends in GDBMSs' usage will be explored and analyzed.

BACKGROUND

In general, there are various definitions of a graph database; De Virgilio et al. defined graph database as a “multigraph $g=(N,E)$, where every node $n \in N$ is associated with a set of pairs $\langle \text{key}, \text{value} \rangle$, and every edge $e \in E$ is associated with a label” (De Virgilio, Maccioni, & Torlone, 2013), whereas He and Singh defined graph database as a set of graphs $D=\{G_1, G_2, \dots, G_m\}$, where graph G is denoted by (V, E) , V being a set of all vertices, and E being a set of all edges (He & Singh, 2006).

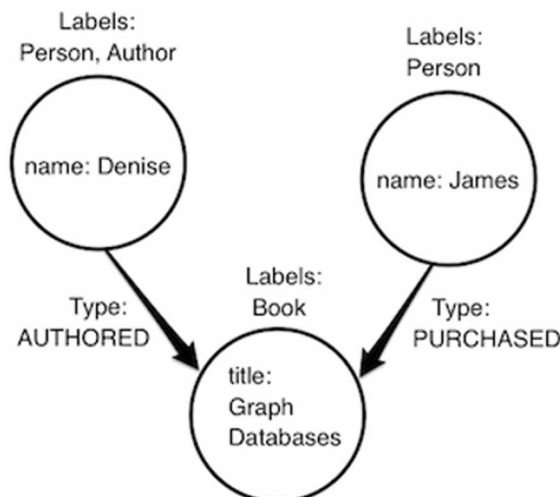
Graph database model can be defined as a data model, in which “data structures for the schema and instances are modeled as graphs or their generalizations, and data manipulation is expressed by graph-oriented operations and type constructors” (Angles & Gutierrez, 2008). Graph database model consists of three components (Angles & Gutierrez, 2008):

- structural component (graph data structures),
- operational component (graph-oriented operators), and
- integrity component (integrity constraints).

Nowadays, the most commonly used graph database model is the property graph data model (Figure 1), which can be defined as a “multigraph data structure, in which graph elements (vertices and edges) can have properties/attributes” (Ciglan, Averbuch, & Hluchy, 2012).

Data stored in a graph database can be queried by using a graph query language, which can provide support for different graph-related operations, such as graph union/intersection/difference, graph filtering, adjacency, path traversal, or pattern matching queries (Kaplan, Abdulla, Brugger, & Kohn, 2007).

Figure 1. Sample property graph data model (Neo4j Inc., 2018)



Hence, a graph database management system (GDBMS) is “a database management system with CRUD (Create, Read, Update, Delete) operations based on property graph data model” (Saha, 2017).

GRAPH DATABASE MANAGEMENT SYSTEMS DEVELOPMENT OVERVIEW

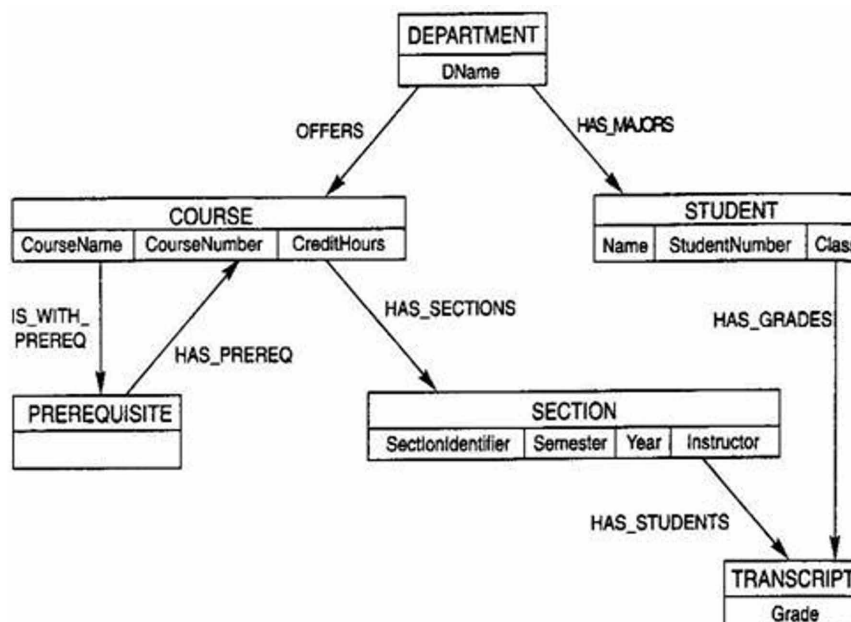
To achieve the objective of this article, i.e., give an overview of past, current and future research trends in the field of graph databases and GDBMSs, a literature review has been conducted to identify relevant published research papers. The analyzed papers were then divided based on their research topic into three categories:

- papers discussing GDBMSs in the past (before 2009),
- papers and topics discussing GDBMSs in the present (2009 – current), and
- papers and topics discussing potential future research trends related to GDBMSs.

Graph Database Management Systems in the Past

The story of storing both entities and their relationships in the database began with hierarchical and network database models, which appeared in 1960s. Between these two, the network database model can be considered as the first “complete” solution for storing connected data in the database at that time due to its ability to store complex and many-to-many relationships (as opposed to only one-to-many relationships available in the hierarchical database model), i.e., in the network database model, a child record can be connected to multiple parent records and vice versa (Maleković, Rabuzin, & Šestak, 2016). The network database model is represented as an inverted tree consisting of records connected with owner-member relationships, which can be implemented as pointers or indexes on the physical level (Harrington, 2002). The most popular network database model used at that time was the one developed in 1971 by CODASYL DBTG¹ (sample model shown in Figure 2), which was then implemented in various DBMSs, and is still used nowadays (Hainaut, 2009).

Figure 2. Sample network data model (Khalil, 2018)



Even though network databases were soon “replaced” by relational databases, nowadays it is still possible to explore their characteristics through Raima Database Manager (RDM) solution developed by Raima Inc., which provides a query language, tools and other utilities to implement a network database model (Maleković et al., 2016).

Graph Database Management Systems in the Present

Research papers published from 2009 until now indicate that graph databases and GDBMSs face challenges common to other database technologies as well.

One of the most important challenges nowadays is storing the growing amount and size of data generated by different applications and businesses. Even though graph databases show good performance in handling such data, it is still necessary to explore how new and trending concepts, mechanisms and technologies can be used to improve their performance and stability in storing and retrieving necessary data in the shortest amount of time (e.g., data compression). The era of cloud solutions and distributed environment represents an additional challenge in scaling and partitioning graph databases without major performance losses.

On a conceptual level, the issue of graph database modeling and introducing some form of graph database schema (less strict than in relational databases) has proved to be of great importance to be standardized and constantly explored, since a correct and accurate graph database model can significantly influence the performance of future graph queries, changes in database structure, etc.

Another great issue quite popular in the research community is the lack of a standardized graph query language. Nowadays, it happens quite often that each modern GDBMS (e.g., Neo4j, TinkerGraph, InfiniteGraph) uses different approaches and query languages to query the underlying graph database, which can be seen in comparison of selected GDBMSs (AllegroGraph, JanusGraph, Neo4j and TinkerGraph) in Table 1. When it comes to graph query languages, there are two important aspects, which need to be considered, which make a query language effective and preferable among its users:

1. the language syntax must remain as simple and understandable as possible for different categories of users ranging from beginners to professionals and domain experts, and
2. the query formulation process must be simplified, but also ensure that the query performance is efficient, effective and stable.

Table 1. Comparison of selected graph DBMSs (Adapted from <https://db-engines.com/en/>)

	AllegroGraph	JanusGraph (former Titan)	Neo4j	TinkerGraph
Primary database model	Graph DBMS RDF store	Graph DBMS	Graph DBMS	Graph DBMS
Data scheme	yes (RDF schema)	yes	schema-free and schema-optional	schema-free
Access methods	RESTful HTTP API SPARQL query language	Java API Tinkerpop computing framework	Cypher query language Java API Neo4j Object Graph Mapping library RESTful HTTP API	Tinkerpop 3 graph computing framework
Transaction concepts	ACID ²	ACID	ACID	no
Foreign keys	no	yes (implemented as relationships in graphs)	yes (implemented as relationships in graphs)	no
License	commercial	open-source	open-source	open-source

The increasing popularity of graph databases (and NoSQL technologies in general) also presents a challenge to convert existing relational database solutions to graph databases to ensure an easy and issue-free transition between the two paradigms. Additionally, in their beginnings, the primary application domain of graph databases and GDBMSs were social networks, fraud detection and recommendation systems. Over the years, the achieved research progress and new GDBMS features have expanded this domain in order to keep up with current technology trends.

Finally, there is a growing tendency to use graph databases for analytical purposes (not only operational) as well, which has led to developing various systems, modules and tools for fast data analysis, as well as different algorithms and methods for data mining in graph databases. Furthermore, the support for implementing RDF (Resource Description Framework) data model in GDBMSs offers an opportunity to handling semantics of data stored in this graph database.

The aforementioned issues and challenges are continually being solved through novel GDBMS architectures, algorithms, frameworks, methods, etc., introduced in both theory and practice, which will be discussed in detail in the following section.

SOLUTIONS AND RECOMMENDATIONS

The topics of graph database modeling and graph query languages have been studied in several research papers. In (Angles et al., 2017), the authors discuss edge-labelled and property graphs as two popular graph data models, different graph querying approaches (graph patterns and navigational expressions), and outline the importance of formalization of graph query languages, such as Cypher, Gremlin or SPARQL. Graph database models and query languages are also discussed in (Soussi, Aufaure, & Baazaoui, 2011).

The process of converting existing relational database solutions to the context of graph databases has been addressed in (Roy-Hubara, Rokach, Shapira, & Shoval, 2017), in which the authors investigate how to create a graph database schema based on an entity-relationship diagram of the application domain. Also, in (Park, Shankar, Park, & Ghosh, 2014), the authors propose a “3NF equivalent graph” transformation to be used for migrating healthcare data from a relational to a graph database.

Modern GDBMSs, such as Neo4j, offer a wide range of features for storing and querying data from the underlying database. However, topics covered in the research community (e.g., graph data partitioning, novel querying approaches, algorithms for query optimization) are often covered only in theory, i.e., rarely implemented in GDBMSs. The most popular graph query language is Cypher, mostly due to its easy-to-learn syntax and pattern matching approach to query processing. According to many users, the biggest drawback for using graph databases is the lack of a standardized query language, which is currently being solved through a so-called “openCypher” project, i.e., an initiative for transforming Cypher language into a standardized graph query language.

Storing billions of nodes and edges is not possible without using some kind of compression techniques and methods. In (Sutrisna, Saleh, & Gozali, 2015), the authors use Graph Algorithm Clustering to compress the graph database, and to produce a lossless and compressed database.

Furthermore, the process of retrieving data from graph database can improve in various segments by using different approaches. S. Das et al. propose a query optimizer for graph databases in (Das, Goyal, & Chakravarthy, 2016). In (Castelltort & Laurent, 2016), the authors discuss how to use in-memory architectures to extract graph database summaries to speed up the graph database querying process. In (Spyropoulos, Vasilakopoulou, & Kotidis, 2016), the authors present “Digree, an experimental middle-ware system that can execute graph pattern matching queries over databases hosting voluminous graph

datasets”. In (Zhang, Gao, Wu, Li, & Gao, 2011), the authors propose a new approach for processing supergraph queries, and a new algorithm for testing subgraph isomorphisms. Another interesting challenge is to find answers to queries over inconsistent graph databases, which is addressed by Barceló and Fontaine in (Barceló & Fontaine, 2017).

In (Yung & Chang, 2017), the authors explore reachability queries, and propose a “How-to-Reach” query to be used when a user does not know, or is unsure how to reach from source to the destination. Similar ideas are presented in (Vasilyeva, Thiele, Bornhövd, & Lehner, 2016), in which the authors propose “Why Empty?” and “Why So Many?” queries, which should explain why a result is empty, or contains too many rows. Queries returning empty answers are also discussed in (Vasilyeva, Thiele, Bornhoevd, & Lehner, 2014); the authors propose the use of top-k differential queries, which should deliver the missing parts of the query graph, so that users have insight into what was missing in the query, and why the query returned no result.

On low-level graph query processing, various algorithms and methods have been introduced to improve the subgraph isomorphism and similarity search processes, which can result in improved query performance (Chodpathumwan, Aleyasen, Termehchy, & Sun, 2016; Kiran & Sivadasan, 2015; Luo, Guan, & Zhou, 2011; Y. Yuan, Wang, Chen, & Wang, 2015).

In addition to query execution performance improvements, several approaches and frameworks have been introduced to simplify the query construction process for end users, such as query auto-completion (Yi, Choi, Bhowmick, & Xu, 2017) and graph exploration (Didimo, Giacche, & Montecchiani, 2015).

Additionally, indexing is always an interesting choice to consider when database performance needs to be improved. In (Azaouzi & Ben Romdhane, 2017), the authors propose the so-called Structural Graph Indexing, called GIRAS, “an indexing feature based on Rare subGraphs (RGs)”, which can be used to decrease the size of the candidate answer set when querying the database. In (D. Yuan & Mitra, 2013), the authors propose a graph index called Lindex, which indexes subgraphs contained within the graph database. In (Jarrar & Deik, 2015), the authors propose a Graph Signature Index suitable for indexing and querying large data graphs. In (Sakr & Al-Naymat, 2010), the authors present different techniques for indexing and querying graph databases.

Another way of improving graph database querying is to use data partitioning. However, graph data partitioning still represents a challenging issue, and only several research papers have been published, which address the issue in question (Barguñó, Muntés-Mulero, Dominguez-Sal, & Valduriez, 2011; Ben Ammar, 2016). Graph data partitioning also represents a basis for building distributed graph databases.

Integrity constraints in graph databases and graph database schema are explored by J. Pokorný et al. in (Pokorný, Valenta, & Kovačič, 2017), in which the authors discuss the level of integrity constraints and schema support in Neo4j GDBMS. In recent years, various graph query languages have been proposed in addition to Cypher, Gremlin and SPARQL graph query languages, which introduce improvements in graph database querying in different segments; a detailed overview was published by P. Wood in (Wood, 2012). For instance, in (Chavarria-Miranda, Castellana, Morari, Haglin, & Feo, 2016), the authors propose a new language called GraQL for attributed graph databases, whereas in (He & Singh, 2008), the authors propose GraphQL query language based on the concept of graph patterns and graphs as the basic unit of information. Additionally, transaction support in graph databases is discussed in (Koloniari & Pitoura, 2016).

An interesting change in GDBMS usage can be noticed in GDBMS market share. Figure 1 shows the ranking of currently top 10 mostly used GDBMSs and their underlying database model according to DB-Engines knowledge base. Based on the figure, it can be observed that among listed DBMSs only two of them are based “purely” on a graph database model (Neo4j and Giraph), whereas the rest

of DBMSs support other database models as well (e.g., documents, key-value pairs, etc.). Therefore, “pure” GDBMSs are slowly being “overtaken” by multi-model DBMSs, which can store data in other database models in addition to graph database model as well.

Figure 3. Top 10 GDBMS ranking (DB-Engines, 2018)

Rank			DBMS	Database Model	Score		
Jun 2018	May 2018	Jun 2017			Jun 2018	May 2018	Jun 2017
1.	1.	1.	Neo4j +	Graph DBMS	41.97	+1.39	+4.10
2.	2.	2.	Microsoft Azure Cosmos DB +	Multi-model ?	19.20	+1.66	+12.82
3.	3.		Datastax Enterprise +	Multi-model ?	7.25	-0.12	
4.	4.	↓ 3.	OrientDB +	Multi-model ?	5.35	+0.10	-0.46
5.	5.	5.	ArangoDB	Multi-model ?	3.52	-0.18	+0.44
6.	6.	6.	Virtuoso	Multi-model ?	1.78	-0.01	-0.22
7.	7.	7.	Giraph	Graph DBMS	0.97	-0.01	-0.08
8.	8.		Amazon Neptune	Multi-model ?	0.67	-0.03	
9.	9.	↓ 8.	AllegroGraph +	Multi-model ?	0.55	-0.03	-0.05
10.	10.	↓ 9.	Stardog	Multi-model ?	0.50	-0.01	-0.04

Currently, many published research papers focus on expanding the applicability of graph database technology to various application domains, such as genetics and medicine, networking, tourism, power systems, etc., which proves that the technology will continue to spread, and that the popularity of graph databases will continually grow in years to come.

FUTURE RESEARCH DIRECTIONS

Even though the popularity of graph databases is constantly increasing, there are still topics and issues that need to be solved. Nowadays, the increasing amount of papers and researched topics indicate several research and development trends in the field of GDBMSs and graph databases, such as semantic graph databases, graph mining, graph data warehouses and graph cubes (Liu & Vitolo, 2013), graph query language standardization initiatives, etc.

For instance, in (Kostylev, Reutter, & Vrgoč, 2016), the authors discuss how to use XPath query language over graphs, whereas in (Lian, Chen, & Wang, 2016), the authors discuss different methods for “quality-aware subgraph matching over inconsistent probabilistic RDF graphs (QA-gMatch)”, which goal is to retrieve subgraphs isomorphic to a given query graph from inconsistent probabilistic RDF graphs.

Additionally, there has been a lot of research focusing on using graph concepts and graph databases for data mining purposes, e.g., frequent subgraph mining (Wang et al., 2016), but also for object, image (Lampoltshammer & Wiegand, 2015) and handwriting recognition (Stauffer, Fischer, & Riesen, 2016).

According to (Green, 2018), a new initiative has been undertaken recently to develop an industry-standardized property graph query language called GQL (Graph Query Language). GQL will combine properties of PGQL (Property Graph Query Language), G-CORE (graph query language used among

researchers) and the openCypher language mentioned in the previous section, and will be used across different GDBMSs (as its counterpart SQL³ in relational DBMSs).

As highlighted in the previous section, current industry trends show, besides Neo4j as the most popular “pure” GDBMS, which is constantly being improved and extended with new features, graph features can also be used in multi-model DBMSs, such as OrientDB or Cosmos DB. Furthermore, the existing relational DBMSs, such as Oracle and Microsoft SQL Server, are starting to extend their support for graph databases as well. This initiative will certainly present an important challenge for Neo4j and other GDBMSs in the future.

CONCLUSION

In this article, an overview of past, current and potential future research and industry trends in the field of GDBMSs and graph databases has been presented. The lack of a standardized graph query language has been identified as currently their major drawback, but research papers published in recent years, as well as current industry initiatives for developing a standardized graph query language (openCypher, GQL) indicate that this problem will soon be resolved. For years, the most widely used GDBMS has been Neo4j. However, recent trends show that other existing DBMSs, which do not support graph data model natively, are continually introducing new features for managing graph-like data (e.g., Oracle). Therefore, it can be presumed that, in the future, Neo4j will be continually developed as a “pure” GDBMS, while other DBMSs will work on partially supporting graph data model.

REFERENCES

- Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., & Vrgoč, D. (2017). Foundations of Modern Query Languages for Graph Databases. *ACM Computing Surveys*, 50(5), 68. doi:10.1145/3104031
- Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys*, 40(1), 1–39. doi:10.1145/1322432.1322433
- Azaouzi, M., & Ben Romdhane, L. (2017). A Minimal Rare Substructures-Based Model for Graph Database Indexing. In *Intelligent Systems Design and Applications (ISDA 2016)* (Vol. 557, pp. 250–259). doi:10.1007/978-3-319-53480-0_25
- Barceló, P., & Fontaine, G. (2017). On the data complexity of consistent query answering over graph databases. *Journal of Computer and System Sciences*, 88, 164–194. doi:10.1016/j.jcss.2017.03.015
- Barguñó, L., Muntés-Mulero, V., Dominguez-Sal, D., & Valduriez, P. (2011). ParallelGDB: a parallel graph database based on cache specialization. In *Proceedings of the 15th Symposium on International Database Engineering & Applications* (pp. 162–169). Academic Press.
- Ben Ammar, A. (2016). Graph Database Partitioning: A Study. *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*. 10.1109/IISA.2016.7785355
- Castelltort, A., & Laurent, A. (2016). Extracting fuzzy summaries from NoSQL graph databases. In *Advances in Intelligent Systems and Computing* (Vol. 400, pp. 189–200). doi:10.1007/978-3-319-26154-6_15

- Chavarria-Miranda, D., Castellana, V. G., Morari, A., Haglin, D., & Feo, J. (2016). GraQL: A Query Language for High-Performance Attributed Graph Databases. In *2016 IEEE 30th International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (pp. 1453–1462). 10.1109/IPDPSW.2016.216
- Chodpathumwan, Y., Aleyasen, A., Termehchy, A., & Sun, Y. (2016). Towards Representation Independent Similarity Search Over Graph Databases. In *CIKM'16: Proceedings of the 2016 ACM Conference on Information and Knowledge Management* (pp. 2233–2238). 10.1145/2983323.2983673
- Ciglan, M., Averbuch, A., & Hluchy, L. (2012). Benchmarking traversal operations over graph databases. In *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on* (pp. 186–189). 10.1109/ICDEW.2012.47
- Das, S., Goyal, A., & Chakravarthy, S. (2016). Plan Before You Execute: A Cost-Based Query Optimizer for Attributed Graph Databases. In *Big Data Analytics and Knowledge Discovery, DAWAK 2016* (Vol. 9829, pp. 314–328). doi:10.1007/978-3-319-43946-4_21
- DB-Engines. (2018). *DB-Engines Ranking of Graph DBMS*. Retrieved June 27, 2018, from <https://db-engines.com/en/ranking/graph+dbms>
- De Virgilio, R., Maccioni, A., & Torlone, R. (2013). Converting relational to graph databases. In *First International Workshop on Graph Data Management Experiences and Systems* (p. 1). Academic Press.
- Didimo, W., Giacche, F., & Montecchiani, F. (2015). KOJAPH: Visual Definition and Exploration of Patterns in Graph Databases. In *Graph Drawing and Network Visualization, GD 2015* (Vol. 9411, pp. 272–278). doi:10.1007/978-3-319-27261-0_23
- Green, A. (2018). *GQL - initiating an industry standard property graph query language*. Retrieved June 30, 2018, from https://db-engines.com/en/blog_post/78
- Hainaut, J.-L. (2009). Network Data Model. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 1900–1906). Springer US. doi:10.1007/978-0-387-39940-9_246
- Harrington, J. L. (2002). *Relational Database Design Clearly Explained*. Elsevier Science. Retrieved from <https://books.google.hr/books?id=HHmBfmJRSL4C>
- He, H., & Singh, A. K. (2006). Closure-tree: An index structure for graph queries. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on* (p. 38). Academic Press.
- He, H., & Singh, A. K. (2008). Graphs-at-a-time: query language and access methods for graph databases. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 405–418). 10.1145/1376616.1376660
- Jarrar, M., & Deik, A. (2015). The Graph Signature: A Scalable Query Optimization Index for RDF Graph Databases Using Bisimulation and Trace Equivalence Summarization. *International Journal on Semantic Web and Information Systems*, 11(2), 36–65. doi:10.4018/IJSWIS.2015040102
- Kaplan, I. L., Abdulla, G. M., Brugger, S. T., & Kohn, S. R. (2007). *Implementing graph pattern queries on a relational database*. Academic Press.
- Khalil, A. (2018). *Database Concepts*. Retrieved October 29, 2018, from <http://www.cs.aucegypt.edu/~csci253/DBConcepts v2.htm>

- Kiran, P., & Sivadasan, N. (2015). Scalable Graph Similarity Search in Large Graph Databases. In *Proceedings of the 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* (pp. 207–211). 10.1109/RAICS.2015.7488415
- Koloniari, G., & Pitoura, E. (2016). *Transaction Management for Cloud-Based Graph Databases*. doi:10.1007/978-3-319-29919-8_8
- Kostylev, E. V., Reutter, J. L., & Vrgoč, D. (2016). Static Analysis of Navigational XPath over Graph Databases. *Information Processing Letters*, 116(7), 467–474. doi:10.1016/j.ipl.2016.03.006
- Lampoltshammer, T. J., & Wiegand, S. (2015). Improving the Computational Performance of Ontology-Based Classification Using Graph Databases. *Remote Sensing*, 7(7), 9473–9491. doi:10.3390/rs70709473
- Lian, X., Chen, L., & Wang, G. (2016). Quality-Aware Subgraph Matching over Inconsistent Probabilistic Graph Databases. *IEEE Transactions on Knowledge and Data Engineering*, 28(6), 1560–1574. doi:10.1109/TKDE.2016.2518683
- Liu, Y., & Vitolo, T. M. (2013). Graph Data Warehouse: Steps to Integrating Graph Databases into the Traditional Conceptual Structure of a Data Warehouse. In *2013 IEEE International Congress on Big Data* (pp. 433–434). 10.1109/BigData.Congress.2013.72
- Luo, Y., Guan, J., & Zhou, S. (2011). Towards Efficient Subgraph Search in Cloud Computing Environments. In *Database Systems for Advanced Applications, DASFAA 2011* (Vol. 6637, pp. 2–13). doi:10.1007/978-3-642-20244-5_2
- Maleković, M., Rabuzin, K., & Šestak, M. (2016). Graph Databases-are they really so new. *International Journal of Advances in Science Engineering and Technology*, 4(4), 8.
- Nabti, C., & Seba, H. (2016). Subgraph Isomorphism Search in Massive Graph Databases. In *IOTBD: Proceedings of the International Conference on Internet of Things and Big Data* (pp. 204–213). 10.5220/0005875002040213
- Neo4j Inc. (2018). *Getting Started with Neo4j and Ruby*. Retrieved October 29, 2018, from <https://neo4j.com/developer/ruby-course/>
- Park, Y., Shankar, M., Park, B.-H., & Ghosh, J. (2014). Graph Databases for Large-Scale Healthcare Systems: A Framework for Efficient Data Management and Data Services. In *2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW)* (pp. 12–19). 10.1109/ICDEW.2014.6818295
- Pokorný, J., Valenta, M., & Kovačič, J. (2017). Integrity constraints in graph databases. *Procedia Computer Science*, 109(Supplement C), 975–981. doi:10.1016/j.procs.2017.05.456
- Roy-Hubara, N., Rokach, L., Shapira, B., & Shoval, P. (2017). Modeling Graph Database Schema. *IT Professional*, 19(6), 34–43. doi:10.1109/MITP.2017.4241458
- Saha, A. L. (2017). *A Comparative Analysis of Graph Vs Relational Database For Instructional Module Development System*. Arizona State University.
- Sakr, S., & Al-Naymat, G. (2010). Efficient Relational Techniques for Processing Graph Queries. *Journal of Computer Science and Technology*, 25(6), 1237–1255. doi:10.1007/11390-010-9402-5

Sieger, R. (2016, April 12). *Graph Databases - Moving Business to the Future*. Retrieved October 29, 2018, from <https://www.linkedin.com/pulse/graph-databases-moving-business-future-robert-sieger/>

Soussi, R., Aufaure, M.-A., & Baazaoui, H. (2011). Graph Database for Collaborative Communities. In *Community-Built Databases: Research and Development* (pp. 205–234). doi:10.1007/978-3-642-19047-6_9

Spyropoulos, V., Vasilakopoulou, C., & Kotidis, Y. (2016). Digree: A Middleware for a Graph Databases Polystore. In *2016 IEEE International Conference on Big Data (BIG DATA)* (pp. 2580–2589). 10.1109/BigData.2016.7840900

Stauffer, M., Fischer, A., & Riesen, K. (2016). A Novel Graph Database for Handwritten Word Images. In *Structural, Syntactic, and Statistical Pattern Recognition, S+SSPR 2016* (Vol. 10029, pp. 553–563). doi:10.1007/978-3-319-49055-7_49

Sutrisna, G. B. A., Saleh, K. R. W., & Gozali, A. A. (2015). Implementation of GRAC Algorithm (Graph Algorithm Clustering) in Graph Database Compression. In *2015 3rd International Conference on Information and Communication Technology (ICoICT)* (pp. 391–395). Academic Press.

Vasilyeva, E., Thiele, M., Bornhoevd, C., & Lehner, W. (2014). Top-k Differential Queries in Graph Databases. In *Advances in Databases and Information Systems (ADBIS 2014)* (Vol. 8716, pp. 112–125). doi:10.1007/978-3-319-10933-6_9

Vasilyeva, E., Thiele, M., Bornhövd, C., & Lehner, W. (2016). Answering “Why Empty?” and “Why So Many?” queries in graph databases. *Journal of Computer and System Sciences*, 82(1), 3–22. doi:10.1016/j.jcss.2015.06.007

Wang, K., Xie, X., Jin, H., Yuan, P., Lu, F., & Ke, X. (2016). Frequent Subgraph Mining in Graph Databases Based on MapReduce. In *Advances in Services Computing* (Vol. 10065, pp. 464–476). doi:10.1007/978-3-319-49178-3_35

Wood, P. T. (2012). Query languages for graph databases. *SIGMOD Record*, 41(1), 50–60. doi:10.1145/2206869.2206879

Woods, D. (2015). How Graph Databases Supercharge Master Data Management. *Forbes*. Retrieved from <https://www.forbes.com/sites/danwoods/2015/08/26/how-graph-databases-supercharge-master-data-management/#18bf1df025b8>

Yi, P., Choi, B., Bhowmick, S. S., & Xu, J. (2017). AutoG: A visual query autocompletion framework for graph databases. *The VLDB Journal*, 26(3), 347–372. doi:10.1007/00778-017-0454-9

Yuan, D., & Mitra, P. (2013). Lindex: A lattice-based index for graph databases. *The VLDB Journal*, 22(2), 229–252. doi:10.1007/00778-012-0284-8

Yuan, Y., Wang, G., Chen, L., & Wang, H. (2015). Graph similarity search on large uncertain graph databases. *VLDB Journal*, 24(2), 271–296. doi:10.1007/00778-014-0373-y

Yung, D., & Chang, S.-K. (2017). Answering How-to-Reach Query in Big Attributed Graph Databases. In *2017 3rd IEEE International Conference on Big Data Computing Service and Applications (IEEE BIGDATASERVICE 2017)* (pp. 141–148). 10.1109/BigDataService.2017.20

Zhang, S., Gao, X., Wu, W., Li, J., & Gao, H. (2011). Efficient algorithms for supergraph query processing on graph databases. *Journal of Combinatorial Optimization*, 21(2), 159–191. doi:10.1007/10878-009-9221-1

Zheng, W., Zou, L., Lian, X., Wang, D., & Zhao, D. (2015). Efficient Graph Similarity Search Over Large Graph Databases. *IEEE Transactions on Knowledge and Data Engineering*, 27(4), 964–978. doi:10.1109/TKDE.2014.2349924

ADDITIONAL READING

Angles, R. (2012, April). A comparison of current graph database models. In Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on (pp. 171-177). IEEE. 10.1109/ICDEW.2012.31

Angles, R., Arenas, M., Barcelo, P., Boncz, P., Fletcher, G., Gutierrez, C., ... van Rest, O. (2018, May). G-CORE: A core for future graph query languages. In *Proceedings of the 2018 International Conference on Management of Data* (pp. 1421-1432). ACM. 10.1145/3183713.3190654

Beis, S., Papadopoulos, S., & Kompatsiaris, Y. (2015). Benchmarking Graph Databases on the Problem of Community Detection. In Bassiliades, N and Ivanovic, M and KonPopovska, M and Manolopoulos, Y and Palpanas, T and Trajcevski, G and Vakali, A (Ed.), *New Trends in Database and Information Systems II* (Vol. 312, pp. 3–14). doi:10.1007/978-3-319-10518-5_1

Holzschuher, F., & Peinl, R. (2013, March). Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops* (pp. 195-204). ACM. 10.1145/2457317.2457351

Macko, P., Margo, D., & Seltzer, M. (2013, June). Performance introspection of graph databases. In *Proceedings of the 6th International Systems and Storage Conference* (p. 18). ACM.

Pokorny, J. (2015). Graph Databases: Their Power and Limitations. In Saeed, K and Homenda, W (Ed.), *Computer Information Systems and Industrial Management* (Vol. 9339, pp. 58–69). doi:10.1007/978-3-319-24369-6_5

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010). A Comparison of a Graph Database and a Relational Database. In *Proceedings of the 48th Annual Southeast Regional Conference (ACM SE 10)* (pp. 223–228). 10.1145/1900008.1900067

Wang, S.-T., Jin, J., Rivett, P., & Kitazawa, A. (2015). Technical Survey Graph Databases and Applications. *International Journal of Semantic Computing*, 9(4), 523–545. doi:10.1142/S1793351X15500129

KEY TERMS AND DEFINITIONS

Graph Data Partitioning: The process of logically partitioning data stored in a graph database into segments, which can then be stored in different locations.

Graph Data Warehouse: A system used for data analysis, in which a data source is a graph database.

Graph Database Model: A conceptual representation of graph-like data in the form of graph nodes and edges.

Graph Mining: The process of discovering patterns in data stored in a graph database through complex data analysis.

Graph Pattern Matching: The process of finding all subgraphs matching a given graph query pattern in the original graph.

Graph Query Processing: The process of transforming a query written using a given graph query language syntax into a low-level query execution plan.

“Pure” Graph Database Management System: A system, which supports CRUD (Create, Read, Update, Delete) operations on graph data represented in the graph database model.

ENDNOTES

- ¹ CODASYL DBTG - Conference on Data Systems Languages – the Data Base Task Group was responsible for developing DBMS specifications for implementing the network data model (Hainaut, 2009).
- ² ACID (Atomicity, Consistency, Isolation, Durability) – transaction concept often present in relational databases, which aims to ensure database integrity.
- ³ SQL (Standard Query Language) – standardized query language for relational databases.

New Trends in Databases to NonSQL Databases

6

Antonio Sarasa-Cabezuelo

 <https://orcid.org/0000-0003-3698-7954>

Facultad de Informática, Universidad Complutense de Madrid, Spain

INTRODUCTION

In recent decades, the relational model has been the mechanism of persistence of information most used in the computer applications. It provides a mature and stable model to represent the data, and also offers a widely used standard query and manipulation (SQL) language.

1. However, the appearance of the “Big Data” phenomenon has meant a change in the storage and information processing needs. This new context is characterized by: a) enormous amounts of information are available in heterogeneous formats and types, b) information must be processed almost in real time, and c) data models evolve periodically.
2. This requires a greater processing capacity, greater flexibility in terms of the types of information data that can be stored and greater flexibility in the data models that facilitate their adaptation.
3. Relational databases have limitations to respond to these needs in an optimal way given that:
 - The solution to achieve more processing capacity is to use more powerful machines (vertical scaling). It is not economically advantageous.
 - The evolution of the data models implies a modification of the database schemas. This causes problems of maintenance and consistency of previously stored information.
 - The storage of heterogeneous and complex data requires the transformation of information in order to store complex data in basic data types (the only ones that manage relational databases).

For these reasons, some companies such as Google or Amazon decided to create new database models (different from the relational model) that solve the needs raised in the context of Big Data without the limitations of relational databases. These new models are the origin of the so-called NonSQL Databases.

Currently, NonSQL databases have been constituted as an alternative mechanism to the relational model and its use is widely extended.

The main objective of this chapter’s proposal is to introduce the NonSQL databases. Specifically:

- To show the processing and persistence needs of information in the Big Data.
- To describe the main characteristics and the conceptual basis of the NoSQL databases.
- To show how the NoSQL databases solve the needs raised in the Big Data
- To describe the main families of NoSQL databases and the data models on which they are based.

The structure of the chapter is as follows. Section 1 describes the storage needs that appear in the Big Data. Section 2 presents the limitations of the databases to cover the needs of Big Data. Next, the

DOI: 10.4018/978-1-7998-3479-3.ch054

NoSQL databases and their main characteristics are introduced. In the following section, persistence and distribution models are discussed. Next, the CAP theorem is presented. Finally, the trends in the NoSQL world are discussed, and a set of conclusions is established.

BACKGROUND

Big Data is a technological challenge that arises due to the confluence of two phenomena (Chen et al, 2014). On the one hand, there is a rapid increase in the storage capacity of the hardware and a reduction in its cost (in the 1980s it was possible to handle data sizes of 0.02 EB, in 2013 they handled sizes of 4.4 ZB, and in 2020 it is estimated that it will be possible to handle sizes of 44 ZB). And on the other hand, some companies and institutions begin to generate large amounts of data in a very fast way (for example, Google generates around 20 PB of data daily and Facebook 500 TB). The possibility of storing this data by having storage with sufficient capacity makes companies start to exploit the stored data for economic or strategic purposes, and to obtain benefits from the results of the analyzes. Thus, an infinite loop begins in which more and more data are generated, which forces us to develop media with greater storage capacity. When initial storage needs have been met, they are no longer able to meet current needs (since the amounts of data generated increase exponentially), which makes it necessary to create new storage media with greater capacity.

Big Data affects different technical aspects related to the storage and processing of information (Khan et al, 2014):

- The types of data that need to be managed are very different in nature, they require the storage of data that can be structured, semi-structured or simply without structure.
- The quantities of data that need to be processed are increasing, so that the necessary processing capacity also increases very rapidly. In this sense, observe that another requirement is the need for the processing to be done almost in real time.
- The data sources that need to be processed can change at any time, which is why we need sufficiently flexible information organization schemes that adapt to different information structures.
- The new context in terms of data types and number of data, makes it necessary to develop new techniques for analyzing information and interpreting the results thereof.

Thus, the technological areas affected by these new needs are varied. However, this chapter will focus on changes in information storage systems. Next, it will be analyzed the limitations of the relationship databases to cover the new needs of Big Data in terms of storage of information.

During the last decades, the relational model has been the most widespread way of carrying out the persistence of information. However, it presents important limitations to cover the main needs arising in the Big Data (Li et al, 2013). Then, each of them will be analyzed (Jacobs, 2009):

- The problem of the data schema (Zhao et al, 2014). In the relational model, a task prior to the storage of the data consists of creating a schema of the type of information to be stored. In this way, they describe the types of data are admitted. If it necessary to store other types of data then it will have to make changes to the scheme. However, these changes often introduce anomalies in the stored data (for example, in terms of the relational model there will probably be rows with many columns with null values). In this sense, it is said that the information stored in a relational

database is structured because it follows a previously defined scheme. In the field of Big Data, the information that needs to be stored can be very diverse, from structured, semi-structured data or even data without structure. In addition, in general it is not known a priori how the type of data will be. For these reasons, it is not possible to define a fixed structure for the information it must be stored. Thus, the characteristics of the data that needs to be stored in a Big Data environment are incompatible with the need to set a previous scheme of data in the relational model, since its use would make it necessary to make changes in the schema for each new type of data that it would have to store. Besides, this situation would introduce anomalies in the database and the need to do changes in the tables and relationships defined.

- The problem of the difference of data structures (Pokorny, 2015). It is a problem related to the previous one. In the relational model, the data structures used to store information are simple data types such as integers, reals, strings, dates... So in order to store real-world information, it must be decomposed in terms of types of data. simple data. On the other hand, the processing of information from a database is usually done from a program written in some information language. To do this, the data is extracted and stored in the data structures offered in the programming language. These structures are more complex than those managed by a relational database, so it is necessary to make a process of composition of the simple data from the relational model to the complex structures of the programming language. In the same way, when you want to store the information managed in a programming language in a relational database, you must perform the inverse process of decomposing complex structures into simple data. These processes of composition and decomposition suppose a cost in terms of efficiency that can not be ruled out when the amounts of information that are going to have to be transformed are enormous, as it is in the case of Big Data.
- The problem of information processing (Jatana et al, 2012). As noted above, the information processing capabilities required in the Big Data field are changing dynamically in an increasing manner. To increase the processing capacity there are essentially two ways: vertical scaling and horizontal scaling. Vertical scaling involves the acquisition of increasingly powerful machines. Thus, when a situation arises in which the machine that performs the processing is not sufficient for the existing needs, then it is replaced by another more powerful machine. However, horizontal scaling is to achieve the necessary processing capabilities through the collaboration of a set of machines with less processing power than the required jointly collaborate to meet the required processing needs (this set of collaborating machines is called cluster of machines). From the economic point of view, vertical scaling is not a good solution because the processing capacity is directly proportional to the cost of the machine (more capacity means higher cost). Thus, if it is necessary to increase the processing capacity it would be necessary to buy a new machine with greater capabilities and therefore with a significant economic cost. However, in horizontal scaling, when using less powerful machines, the total cost of the necessary machines will be lower. In addition, if it is necessary to increase the processing capacity it would be enough to buy new machines with lower capabilities (and therefore with a lower economic cost compared to vertical scaling) that would be added to the current machine cluster. Therefore, the problem of the increasing need for processing capacity is solved in an optimal way using horizontal scaling. However, this solution is incompatible with the use of relational databases, since they are designed to run in centralized environments on a single machine, and not to run on a set of machines where processing is distributed. The distributed execution of a relational database introduces problems in aspects such as performing operations of type join or on the distribution of data rows of a table.

SOLUTIONS AND RECOMMENDATIONS

Due to the limitations of the relational model (discussed in the previous section) to cover the storage needs of Big Data, some companies such as Google (Big Tables) or Amazon (Dynamo) developed new alternative persistence systems (Kalid et al, 2017) to the relational model (Chen et al, 2014). A common feature was its ability to manage and carry out the processing of huge amounts of data in distributed environments. These alternative systems are the origin of the so-called NoSQL Databases (No SQL), which served as a starting point for the development of other alternative systems. It is complex to define a classification of the different types of NoSQL Databases (Gessert et al, 2017), since many of them share characteristics. However, there is a first level classification that defines several families (Nayak, 2013):

- Key-value data bases such as Riak, Redis, Dynamo or Voldemort (Kalid et al, 2017)
- Document-oriented databases such as MongoDB or CouchDB (Jose et al, 2017)
- Databases based on columns such as Cassandra, Hypertable, HBase or SimpleDB
- Databases of graphs such as Neo4J or Infinite Graph.

There is a set of characteristics shared by most of the NoSQL databases (Han et al, 2011):

- 1) One of the most important features is that they do not use SQL as a query language. In general, all of them have a query language to be able to manage and consult the information they store, but their syntax and the type of possible operations are different from the SQL language of the relational databases. However, there is a case in which the query language has a certain similarity with SQL, such as the CQL language of the Cassandra database (Hsieh et al, 2017)
- 2) In most cases these are open source projects.
- 3) Another feature that most of these types of databases fulfill is that they have been designed to run in environments distributed over a cluster of machines. This feature has influenced the underlying data models used to store the information or solutions used to address the consistency problem. However, some models do not fulfill this characteristic, as is the case of graph-oriented databases that run in centralized environments.
- 4) They do not use a fixed information scheme, so that information can be added freely without being subject to any structure (Gupta et al, 2017). This feature has several implications with respect to the limitations of the relational databases discussed above. On the one hand, the problem of having to previously fix the types of data that can be stored is solved, any type of data compatible with the data model of the database used can be stored. And on the other hand the problem of the cost of the process of composition and decomposition of the data is solved, given that the data structures used in this type of databases are similar to those used in the programming languages. However, two aspects of this flexibility in terms of information schemes must be observed. On the one hand, some tasks performed by database management systems such as correcting the data entered, how to store and retrieve information in the most efficient way, or checking the consistency of the data that are manipulated by the applications that access the database, now they will have to be carried out directly by the programmers (Branch et al, 2017). And on the other hand there is the problem of knowing information about stored data such as the type of data, names of the information fields... In this sense, although there is no explicit schema in the database, if there is an implicit schema in the programs that manage the database. Thus, some aspects such as where each type of data is stored, what are the data formats that can be used, how is the structure of the database..., are described in

the programs that access it. This is a weakness, for example when changes occur in the database, since these changes will have to be transferred to the programs.

With regard to distribution models, two distribution models can be found: replication and sharding (Hecht et al, 2011). Replication consists of having copies of the data managed by the distributed database on multiple machines in the cluster. Within this model, you can find some two variants: the Master-Slave model and the Peer-to-Peer model. The first model is characterized because there is a machine called “Master” that stores the reference copy of the system. Thus, all the deeds and modifications of the data are made on this copy. The rest of the machines called “Slaves” keep copies of the Master data, and they support the data readings that are made on the system. Likewise, the slave machines must be kept synchronized with the Master machine so that any change that occurs in the copying of reference data is reflected in the copies of the slave machines. The second model is characterized in that all machines play the same role, and can receive both readings and data writes, which requires synchronization of all the machines in the cluster every so often in order to maintain the consistency of the data. The advantage of the first model is the reduction of inconsistencies in the information that is updated (since it is done in a single machine) while the second model has the advantage of reducing the overload of data writing and modification operations, given that it is done on any of the machines in the cluster.

On the other hand, sharding (Pokorny, 2013) consists of distributing the managed data in different machines, so that each of them will have a subset of the totality of managed data. This model has advantages in aspects such as the possibility of managing large amounts of data, the possibility of processing a greater number of read and write operations, presenting greater availability before failure situations in the communications network,... However, the cost associated with the complexity of the system is the main disadvantage

These are two orthogonal models, and are generally used in combination. A replicated and distributed system has the advantages of guaranteeing access to data (even if one machine fails, you can always access another one that maintains a copy of the data) and make it easier to scale up the processing capacity when the data increased (Just it is necessary add more machines to the cluster to distribute the new data).

Regarding the consistency of the information (Stonebraker, 2012), it should be noted that one of the characteristics of the NoSQL databases is that this aspect is no longer a task of the database management system, and becomes a task of the programmer. In this sense, the ACID model of transactions is not guaranteed in this context. Thus, different types of conflicts can be found regarding the consistency of the information: writing-writing conflict, reading-writing conflict and reading conflict. The first case occurs when you try to perform two updates of the information at the same time. In this case, the problem of the order in which the updates are made is raised since inevitably one of them will overwrite the other. The second type of conflict occurs when you try to perform a reading at the same time you try to write on the same data. Depending on the order in which the operations are carried out, an inconsistency in the read data may occur. The third type of conflict occurs in systems where data replication has been used. In these systems, there is a period of time called “window of inconsistency” during which some slave machines are in an inconsistent state with respect to the Master machine where the update occurred (Corbellini et al, 2017). In this situation, if a reading occurs on a slave machine, the information read may be inconsistent with respect to the updated information. Observe that it is possible to design a system that avoids inconsistencies, but will have the cost of sacrificing other characteristics. Likewise, we must bear in mind that each domain has a different tolerance to inconsistency (it is not the same to allow an inconsistency in an online newspaper than in a web page that is dedicated to the purchase and

sale of shares of the Stock Exchange), and that different levels of consistency could be used based on the type of information that is managed or the type of update to be made.

There is a theorem called “CAP Theorem” applicable to distributed systems that run in a cluster of machines such as the NoSQL databases (Li et al, 2013). It establishes that given a distributed system if one considers the properties of consistency, availability (access to a machine of a cluster to read and write data) and tolerance to partitioning (to support breaks of communication between the elements of the system), then it is only possible guarantee 2 of them at a time. Thus, the theorem states that a system can suffer partitions and consistency must be compensated to achieve availability. In this sense, the system will not be completely consistent or available, but it will have a reasonable combination (depending on the domain it is possible to maintain certain inconsistencies during a certain period of time allowing increasing the availability and performance of the system).

FUTURE RESEARCH DIRECTIONS

NoSQL databases have become an alternative to the widely accepted relational model. There are two trends currently regarding its use (Davoudian et al, 2018):

- a. Polyglot persistence (Assunção et al, 2015). NoSQL databases should not be used in all cases of data persistence. In many situations, the information storage needs require mixed solutions where a relational database is combined that will be in charge of managing structured data with a NoSQL database that will be responsible for managing semi-structured or unstructured data (Haseeb et al, 2017). In these cases, the communication between both systems will be done through a web services API that hides the way to access and retrieve the information. The advantages of these solutions is the efficient adaptation to the needs of each type of information that must be managed by the system. It also allows to scale and adapt each type of database in a transparent manner without affecting the rest of the solutions.
- b. The NewSQL databases (Kumar, 2014). Another trend in the NoSQL world are the so-called NewSQL databases. It is a set of tools (Stonebraker, 2010) that maintain both the use of the relational model, the SQL query language to manage the stored data and the ACID model but achieving the same scalable performance as the NoSQL databases to process reading transactions -writing. A common feature (Paylo et al, 2016) to all of them is that they have been designed to run in a distributed cluster, so that each node has a subset of the managed data. Regarding the processing of queries, there are different approaches. One of the most widespread techniques (Niazi et al, 2017) is to perform transactions and queries to the cluster machines that contain the necessary data, so that the queries are divided into fragments.

CONCLUSION

In this chapter it has presented a new type of databases called NoSQL that emerge as an alternative to the relational model. The reasons for its appearance are found in the limitations of the relational databases to cover the persistence needs that appear in the Big Data (essentially the need to process data of any kind, in huge quantities, in a fast way and in distributed environments). These persistence systems have some common characteristics, such as not using SQL as a query language, not using a fixed schema to

represent the structure of the information they store and, in most cases, running in a cluster of distributed machines. These characteristics have a cost in terms of the consistency of the information since they do not guarantee the ACID properties, and also many of the tasks previously performed by the database management systems in terms of data validation, correction of the information that is stored..., now they must be done by the programmer. Precisely with respect to the consistency of the information, different types of inconsistency arise that must be taken into account in the design of the persistence solutions (and that will depend on the domains in which it will be used): writing-writing inconsistency, reading-writing inconsistency or reading inconsistency. Another aspect that arises in this context are the different ways of distributing the data, which can be done in two different ways: replication and sharding. Each of these solutions has its own characteristics, although generally there is a tendency to implement combined solutions of distributed data replication. Finally, it is mentioned that there are two trends mainly in the NoSQL world: NewSQL databases (Pavlo et al, 2016) and mixed persistence solutions. In the first case, they are databases that maintain the relational model but allow its distributed execution as well as other similar characteristics of the NoSQL databases. And mixed persistence solutions propose using the best solution, whether relational or NoSQL, for each type of data that must be managed by a system.

REFERENCES

- Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79, 3–15. doi:10.1016/j.jpdc.2014.08.003
- Branch, S., Branch, B., & Boroujen, I. (2017). A novel method for evaluation of nosql databases: A case study of cassandra and redis. *Journal of Theoretical and Applied Information Technology*, 95(6).
- Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314–347. doi:10.1016/j.ins.2014.01.015
- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171–209. doi:10.1007/11036-013-0489-0
- Corbellini, A., Mateos, C., Zunino, A., Godoy, D., & Schiaffino, S. (2017). Persisting big-data: The NoSQL landscape. *Information Systems*, 63, 1–23. doi:10.1016/j.is.2016.07.009
- Davoudian, A., Chen, L., & Liu, M. (2018). A Survey on NoSQL Stores. [CSUR]. *ACM Computing Surveys*, 51(2), 40. doi:10.1145/3158661
- Gessert, F., Wingerath, W., Friedrich, S., & Ritter, N. (2017). NoSQL database systems: A survey and decision guidance. *Computer Science-Research and Development*, 32(3-4), 353–365. doi:10.1007/00450-016-0334-3
- Gupta, S., & Narsimha, G. (2017). Efficient query analysis and performance evaluation of the NoSQL data store for bigdata. In *Proceedings of the First International Conference on Computational Intelligence and Informatics* (pp. 549-558). Springer. 10.1007/978-981-10-2471-9_53
- Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on* (pp. 363-366). IEEE.

- Haseeb, A., & Pattun, G. (2017). A review on NoSQL: Applications and challenges. *International Journal of Advanced Research in Computer Science*, 8(1).
- Hecht, R., & Jablonski, S. (2011). NoSQL evaluation: A use case oriented survey. In *Cloud and Service Computing (CSC), 2011 International Conference on* (pp. 336-341). IEEE. 10.1109/CSC.2011.6138544
- Hsieh, M. J., Ho, L. Y., Wu, J. J., & Liu, P. (2017). Data partition optimisation for column-family NoSQL databases. *International Journal of Big Data Intelligence*, 4(4), 263–275. doi:10.1504/IJBDI.2017.086962
- Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8), 36–44. doi:10.1145/1536616.1536632
- Jatana, N., Puri, S., Ahuja, M., Kathuria, I., & Gosain, D. (2012). A survey and comparison of relational and non-relational database. *International Journal of Engineering Research & Technology*, 1(6).
- Jose, B., & Abraham, S. (2017). Exploring the merits of nosql: A study based on mongodb. In *Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on* (pp. 266-271). IEEE. 10.1109/NETACT.2017.8076778
- Kalid, S., Syed, A., Mohammad, A., & Halgamuge, M. N. (2017). Big-data NoSQL databases: A comparison and analysis of “Big-Table”, “DynamoDB”, and “Cassandra”. In *Big Data Analysis (ICBDA), 2017 IEEE 2nd International Conference on* (pp. 89-93). IEEE.
- Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, M., Kamaleldin, W., ... Gani, A. (2014). Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal*. PMID:25136682
- Kumar, R., Gupta, N., Maharwal, H., Charu, S., & Yadav, K. (2014). Critical Analysis of Database Management Using NewSQL. *International Journal of Computer Science and Mobile Computing*, 3, 434–438.
- Li, Y., & Manoharan, S. (2013). A performance comparison of SQL and NoSQL databases. In *Communications, computers and signal processing (PACRIM), 2013 IEEE pacific rim conference on* (pp. 15-19). IEEE. doi:10.1109/PACRIM.2013.6625441
- Nayak, A., Poriya, A., & Poojary, D. (2013). Type of NOSQL databases and its comparison with relational databases. *International Journal of Applied Information Systems*, 5(4), 16–19.
- Niazi, S., Ismail, M., Haridi, S., Dowling, J., Grohsschmiedt, S., & Ronström, M. (2017). HopsFS: Scaling Hierarchical File System Metadata Using NewSQL Databases. In *FAST* (pp. 89-104). Academic Press.
- Pavlo, A., & Aslett, M. (2016). What’s really new with NewSQL? *SIGMOD Record*, 45(2), 45–55. doi:10.1145/3003665.3003674
- Pokorny, J. (2013). NoSQL databases: A step to database scalability in web environment. *International Journal of Web Information Systems*, 9(1), 69–82. doi:10.1108/17440081311316398
- Pokorny, J. (2015). Database technologies in the world of big data. In *Proceedings of the 16th International Conference on Computer Systems and Technologies* (pp. 1-12). ACM. 10.1145/2812428.2812429
- Stonebraker, M. (2010). SQL databases v. NoSQL databases. *Communications of the ACM*, 53(4), 10–11. doi:10.1145/1721654.1721659
- Stonebraker, M. (2012). Newsql: An alternative to nosql and old sql for new oltp apps. *Communications of the ACM*. Retrieved, 07-06.

Zhao, G., Lin, Q., Li, L., & Li, Z. (2014). Schema conversion model of sql database to nosql. In *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2014 Ninth International Conference on* (pp. 355-362). IEEE.

ADDITIONAL READING

Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education.

Tiwari, S. (2011). *Professional NoSQL*. John Wiley & Sons.

Vaish, G. (2013). *Getting started with NoSQL*. Packt Publishing Ltd.

Warden, P. (2011). *Big data glossary*. O'Reilly Media, Inc.

KEY TERMS AND DEFINITIONS

Big Data: It is a phenomenon that involves a set of tools, technologies and techniques that aim to process huge amounts of data of all types in order to exploit the information contained and obtain some kind of advantage.

Consistency: It refers to the mechanism that ensures that the data that is retrieved from a database is valid and is updated.

NoSQL Database: It is a database that is not designed using the relational model as a basis. It presents, among other features, that it does not use the SQL language as a query language.

Relational Database: It is a database that is designed using the relational model as a basis. Its main characteristic is that it organizes the information using tables as a storage unit. Likewise, it uses SQL as a query language.

Replication: It consists in making a copy of the data in more than one machine of a distributed system in order to guarantee that the data is not lost in the case of the fall of any of the machines.

Scaling: Refers to the way you can increase the processing capacity of a system. There are two types: vertical scaling and horizontal scaling. In the first case, it consists of acquiring a machine with better features than the previous one, while in the second case, it consists of creating a cluster of machines that collaborate with each other, so that the processing powers of all the machines are added.

Sharding: It is an organization of information that consists of distributing data among a set of machines so that machines can define roles such as a machine acting as a server and the rest of customers or that all machines act with the same roles.

SQL Language: It is a declarative language that is used in relational databases to make queries.

Theorem CAP: It is a theorem that establishes a set of properties about the distributed systems with respect to the availability, consistency of the data or the partitioning of the system.

EDRC: An Early Data Lending–Based Real– Time Commit Protocol

Sarvesh Pandey

 <https://orcid.org/0000-0002-3014-9792>

MMM University of Technology, India

Udai Shanker

 <https://orcid.org/0000-0002-4083-7046>

MMM University of Technology, India

INTRODUCTION

The advanced Distributed Database System (DDBS), with the integration of ‘real-time’ attribute to each transaction, makes sure that results are produced within a stipulated deadline while maintaining the consistency. Such systems are broadly studied under the umbrella of the research area named Distributed Real Time Database System (DRTDBS). In simple, they are time constrained DDBSs (Pandey & Shanker, 2018a). In the DRTDBS, the distributed real-time transaction (DRTT) is invoked to perform changes at multiple sites atomically. Here, the correctness of the result depends on two things: logical computation performed, and the time when the result is produced. Even if the result produced is functionally right, it may lead to tragic repercussions, be unusable, or has less value if it is not produced in time (Shanker, Misra, & Sarje, 2001).

Based on the consequences of their deadline misses, DRTTs may be categorized as soft, firm and hard. The soft DRTT is not killed/aborted in case of its deadline miss because the result has some value (obviously degrading) even after the deadline miss. The firm DRTT is killed in case it misses its deadline because its outcome has no value after the deadline miss. In addition, allowing the execution of firm DRTT after its deadline expiry may also lead to deadline expiration of other concurrently executing firm DRTTs requiring access to resources already locked by it. The hard DRTT must be finished before its deadline; otherwise, it may lead to a potentially catastrophic consequence.

Scheduling concurrent execution of DRTTs is extremely complex as the goal of database management algorithms is not only to maintain database consistency but also to ensure that timing constraints are met. Issues such as data conflicts, site failures, communication delays, interaction through the underlying operating system and I/O subsystem are major obstacles in meeting the deadlines of concurrently executing DRTTs (Shanker U., 2008). Among all these, the data conflict problem is a key factor that adversely affects system performance; it can be categorized as Execute-Execute Conflict and Execute-Commit Conflict. The Execute-Execute conflict may occur amongst executing transactions. A considerable research investigation has been carried out to resolve this conflict. For detailed knowledge about the Execute-Execute conflict and strategies proposed for its resolution, the readers may go through (Shanker, Misra, & Sarje, 2008). The Execute-Commit conflict may occur between executing and committing transactions. It occurs only if one of the transactions involved in data conflict is in a PREPARED state.

DOI: 10.4018/978-1-7998-3479-3.ch055

The real-time concurrency control protocol resolves execute-execute conflict while the real-time commit protocol, an optimistic implementation of 2-Phase Commit (2PC) protocol with relaxed ACID property, resolves execute-commit conflict.

To overcome execute-commit conflicts, the lender-borrower approach is commonly used in next generation commit protocols — which are not only of optimistic nature, but also intended to satisfy real-time constraints. Such real-time commit protocols permit prepared lender cohorts to lend their modified and uncommitted data items to executing borrower cohorts in case of occurrence of Execute-Commit conflict between them. This reduces data inaccessibility. Thus, they improve DRTDBS performance. They, however, usually do not permit the borrower cohort to send WORKDONE/PREPARED message, and thus, increase the borrowing transactions' commit time. Hence, designing efficient real-time commit protocol is a vital issue as it largely affects the DRTDBS performance.

BACKGROUND

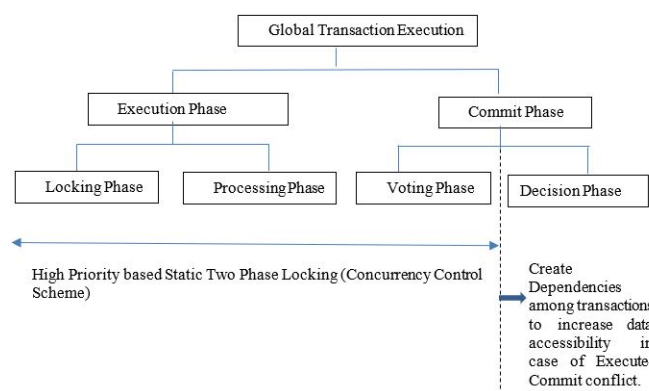
The research on developing an efficient real-time variant of the classical 2PC is still an open question in the study of the DRTDBS (Pandey & Shanker, 2016). Ramesh Gupta et al. first proposed a real-time variant of classical 2PC protocol named OPT (Gupta, Haritsa, Ramamritham, & Seshadri, 1996); this protocol is specifically designed to fulfill the consistency and deadline requirements of the DRTDBS by reducing the negative impact of Execute-Commit conflict and inherent priority inversion problem. The OPT protocol permits a high priority cohort to borrow/access the uncommitted data item(s) held by prepared low priority cohort. Such lending by a prepared cohort creates a dependency between conflicting DRTTs. The fate of borrower DRTT depends on the final outcomes of DRTTs from whom it had borrowed the data items — if final outcomes of all the lender DRTTs are 'commit', then only the borrower DRTT can start its commit processing. Thus, the OPT makes the borrower to be blocked till lenders complete their execution; this optimism is advantageous only when lenders successfully complete their executions. While the policy of using uncommitted data items may result in the chain of dependencies responsible for the cascading aborts, the OPT protocol restricts the length of chain to only one by not allowing the borrowers to simultaneously become lenders. Therefore, it doesn't suffer from a cascading abort problem. It provides a considerable performance improvement over 2PC.

Moreover, two variants of the OPT is also suggested: Healthy-OPT and Shadow-OPT (Gupta, Haritsa, & Ramamritham, 1997). In the Healthy-OPT, every executing transaction is assigned a health factor (HF), and a transaction can become lender only when its HF value is greater than or equal to some threshold value. Obviously, the performance of the system heavily varies with change in the chosen threshold value. In the Shadow-OPT, the cohort originates a shadow (replica of itself) at the instance it borrows some uncommitted data item. Moreover, the original version of the borrower cohort continues its execution as usual even after borrowing uncommitted data item while the shadow cohort is blocked right after being forked off as a separate cohort. In case lenders successfully commit, the shadow cohort is discarded since optimistic borrowing done its job well. Otherwise, if any of the lenders (who have a dependency with this cohort) aborts, the shadow cohort is activated, and the original cohort is aborted. Thus, the Shadow-OPT saves the borrower cohort from restarting its execution in case of an unsuccessful lending-borrowing event and provides a feature so that the borrower cohort can resume its processing from the point the borrowing is made. The detailed simulation results showed that the Healthy-OPT performs reasonably fairer than the Shadow-OPT.

Later, the ‘Permits Reading of Modified Prepared Data for Timeliness’ (PROMPT) protocol is proposed; it is an extension/ integration of the work done in the OPT protocol and its variants (Haritsa, Ramamritham, & Gupta, 2000). The PROMPT provides features like controlled optimistic access to uncommitted data, active abort, silent kill, and healthy lending. Though the PROMPT protocol remained the standard for a long time to assess the implementation-performance of other real-time commit protocols that came after it, several shortcomings of this protocol are raised by researchers with time. Some key developments are discussed below.

To reduce data inaccessibility and priority inversion duration, the Double Space Commit (2SC) protocol exhibits two properties (Qin & Liu, 2003). The 2SC protocol, unlike PROMPT, has two main features: (i). permits a non-healthy prepared transaction to lend the data, that is earlier locked by it in shared mode, to the transactions that lead to commit dependency with it. (ii). in case of the abort of the prepared lender transaction, transactions from its abort dependency set are aborted. Such abort does not affect the execution of transactions that are in the commit dependency set of the prepared lender transaction. The 2SC protocol performs fairer than the PROMPT protocol. The ‘Static two-phase locking and high priority based, Write-update type, Ideal for Fast and Timeliness commit protocol’ (SWIFT) protocol (Shanker, Misra, & Sarje, 2006) analyzed all type of dependencies that may happen during execute-commit conflicts. The SWIFT divides the execution phase of the cohort into two phases: (i). locking and (ii). processing phase. Here, the cohort sends an early message named as WORKSTARTED message just after the completion of the locking phase. This leads to a situation in which the WORKDONE message need not be sent after the completion of the processing phase of executing cohort. The changed scenario leads to an improved system performance since the delay involved in sending the WORKDONE message is eliminated; it became possible through an innovative means of overlapping the send event of WORKSTARTED message and the processing time involved in executing a cohort. The SWIFT protocol enhanced the system performance as compared to PROMPT and 2SC protocol by reducing transaction miss percentage up to 10% (Shanker, Misra, & Sarje, 2009). The working of the SWIFT protocol may be clearly understood using Figure 1.

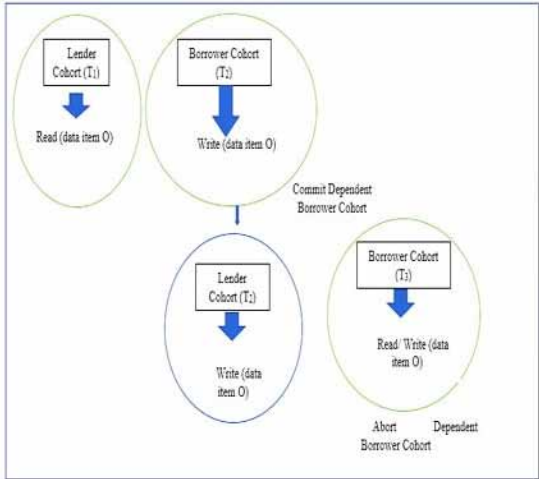
Figure 1. Level of Data Accessibility in SWIFT Protocol



The Extended SWIFT protocol (Aakash, Gaurav, Gupta, & Kumar, 2015), proposed by Ankit Aakash et al., discussed the database inconsistency problem resulting due to Read-Read access in the DRTDBS. It also uses the Lender-Borrower approach to deal with the priority inversion. ACTIVE protocol (Shanker, Agarwal, Tiwari, Goel, & Srivastava, 2010) categorizes borrowers as following; (i). Commit dependent

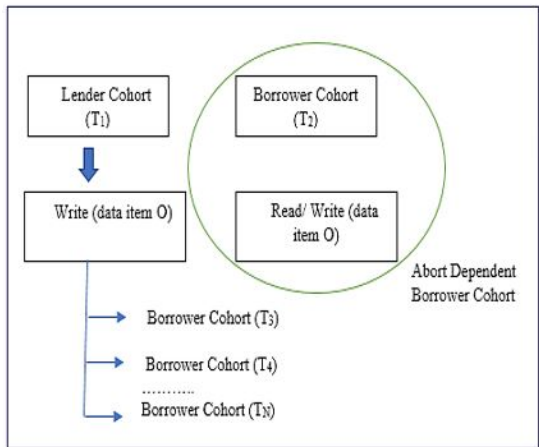
borrower and (ii). Abort dependent borrower. To reduce the data inaccessibility, the commit dependent borrowers are further allowed to lend their uncommitted data. Here, a metric named ‘borrowing factor’ is formulated and computed for each borrower distinctly. This ensures that the lock can be granted to a second-step abort dependent borrower fruitfully by a first-step commit dependent borrower. The working of the ACTIVE protocol may be clearly understood using Figure 2.

Figure 2. Level of Data Accessibility in ACTIVE Protocol



In PERDURABLE protocol (Shanker, Vidyareddi, & Shukla, 2012), a lender writer cohort that has already provided access to uncommitted data to an abort dependent borrower cohort can provide access to the same data in shared mode to other borrowers. It further improves the data item availability by allowing all the locks, that were not required during the processing phase, to be released before the start of the commit phase. The working of the PERDURABLE protocol may be clearly understood using Figure 3.

Figure 3. Level of Data Accessibility in PERDURABLE Protocol



As a part of one more advancement in this direction, an Improved Data Lending based Distributed Real-Time Commit (IDRC) protocol (Pandey & Shanker, 2017a) has been developed; it lessens the effect of the Executing-Committing conflict on the DRTDBS performance by providing the better data accessibility in comparison with earlier protocols. It does so by allowing the commit-dependent first-step borrower (who has already provided access of uncommitted data item to the second-step abort dependent borrower) to lend the same data item in shared mode to numerous other second-step borrower cohorts.

Most of the above-discussed protocols, extending the PROMPT protocol, specifically reduce data inaccessibility to achieve their common goal. As a successor of them, the proposed Early Data-Lending based Real-Time Commit (EDRC) protocol provides a feature to start the lending process just after the completion of the data processing task during the processing phase obviously after locking phase. As a result of allowing lending to happen even before the voting phase of a cohort, the EDRC reduces the transaction deadline miss percentage.

EDRC: A PROPOSED REAL-TIME COMMIT PROTOCOL

Several real-time conflict resolution strategies (concurrency control protocols and commit protocols) have been proposed till date for resolution of data conflicts among concurrently executing DRTTs. These strategies heavily affect DRTDBS performance. Lower the wastage of resources as a consequence of aborts, higher will be the chance of completion of the concurrently executing DRTTs. But there is a need to ensure that the protocols should be well-concentrated toward valuing the priority order instead of reducing wastage of resources or increasing throughput. Hence, the primary DRTDBS performance objective is to reduce the number of DRTTs missing their deadlines; obviously the consistent and atomic commitment of DRTTs must be guaranteed as well. The above objective can be achieved by means of increasing data accessibility. However, blocking the DRTTs — the main source of data inaccessibility — is necessary to ensure the database consistency. The data inaccessibility problem may occur when a set of requesting DRTTs gets blocked from accessing an uncommitted data item which is currently being held by a prepared cohort of other DRTT. So, data conflicts among concurrently executing DRTTs may give birth to the data inaccessibility.

To address the above problems more efficiently, an Early Data-Lending based Real-Time Commit (EDRC) protocol is proposed which not only categorizes the data conflicts among DRTTs but also provides different ways to handle these newly described data conflicts. The EDRC protocol optimistically reduces wastage of resources by employing the cooperative lender-borrower strategy through the early start of data lending during the concurrent execution of DRTTs, and in turn reduces the number of DRTTs missing their deadlines. Another advantage with EDRC protocol is that the positive features of the existing lender-borrower based real-time commit protocols can be easily integrated with it. The proposed protocol is an extension of the SWIFT protocol (Shanker, Misra, & Sarje, 2006). It divides the execution phase into 2 phases: the locking phase and processing phase.

The cohorts lock the data items required by them during their locking phase only. After completion of the locking phase, cohorts send WORKSTARTED messages to their respective coordinators; this event occurs at the beginning of the processing phase. After receiving the WORKSTARTED message from its cohorts, the coordinator sends PREPARE message to all its cohorts at time t . This time is computed as follows.

$$t = \text{MAX} \{t_i + \text{ProcessingTime}_i\} - t_{CD}$$

where t_i = arrival time of cohort_{*i*}'s WORKSTARTED message, $ProcessingTime_i$ = Processing time involved in executing cohort_{*i*}, and t_{CD} = Comm. Delay.

To improve the data accessibility, the EDRC protocol re-categorizes the data conflict occurring among DRTTs as following: locking-locking conflicts, locking-processing conflicts, and locking-committing conflicts. The locking-locking conflict and locking-processing conflict both may occur between executing transactions; hence, they are a result of the further classification of execute-execute conflict (Pandey & Shanker, Transaction Scheduling Protocols For Controlling Priority Inversion: A Review, 2020), a widely studied data conflict in literature. The proposed protocol, in contrary to the SWIFT, handles the locking- locking and locking-processing conflicts differently. The locking-locking conflict is handled in the same way as execute-execute conflict handled by previously developed protocols simply by aborting the DRTT with low priority in case of priority inversion.

The handling of the locking-processing conflict, however, makes the EDRC protocol better than SWIFT. In case of the locking-processing conflict, there is a chance that the executing cohort of a low priority DRTT may complete its execution without impairing some other executing cohort(s) of high priority DRTT(s) that have conflicting data access requirements. This prospect is exploited by initiating the lending process at an early stage, given that this executing cohort which belongs to the low priority DRTT has not only sent the WORKSTARTED message to its coordinator but also completed its data processing task before the occurrence of the above locking-processing data conflict.

Let us consider the case where some cohort of high priority DRTT requests access to the data item in its locking phase which is already being locked by executing cohort of some other low priority DRTT. This conflict may be named as locking-processing conflict. For resolving execute-execute conflicts, previous real-time commit protocols had a provision to simply abort the lock holding DRTT in case its priority is lower than that of lock requesting DRTT. In contrary to this, the EDRC protocol provides a mechanism to start the lending process among DRTTs just after the completion of the data processing task by a cohort in case of data conflict with this cohort. To efficiently handle the data conflict among executing transactions, the lender-borrower approach is utilized as below.

1. When the request for accessing the data item which is already modified by an executing cohort of some DRTT is generated by another executing cohort of some other DRTT, it is checked whether the lock holding cohort has completed its data processing task or not.
2. If the lock holding cohort has completed its data processing task before such conflicting lock request then it can immediately provide dirty data item to the borrower DRTT in a dependent manner as opposed to existing real-time commit protocols; otherwise, if lock holding cohort's data processing task is not completed, then such conflict is resolved using the S2PL-HP protocol.
3. By doing this, the priority inversion is not only optimistically handled but the resource wastage is also avoided to some extent. It results in a reduction in overall DRTT execution time, an increase in data accessibility and the most important minimization in the transaction deadline miss percentage.

Classification of Transactional Dependency

Transactional dependency may arise among a set of DRTTs requiring access to a conflicting data item without being blocked. Such dependencies restrict the commit order of conflicting DRTTs. There can be two types of transactional dependencies: (i). transactional commit dependency (TCD) and (ii). transactional abort dependency (TAD). When the DRTT, ' T_i ' requests a write lock on a data read by other DRTT ' T ', it leads to a TCD from T_i to T represented by T_i TCD T . Here, the younger T_i is not enabled

to commit until the older T has committed. When the DRTT, ' T_i ' performs a read/write operation on an uncommitted data modified/ updated by DRTT 'T', a TAD is established from T_i to T represented by T_i TAD T. Here, the younger T_i must commit after the older T and the abort of T forces T_i to abort. Each site with executing DRTTs maintains a table with the following information about local active cohorts. For the ACID compliant DRTT execution, the above two transactional dependencies need to be controlled logically.

Computation of Health Factor

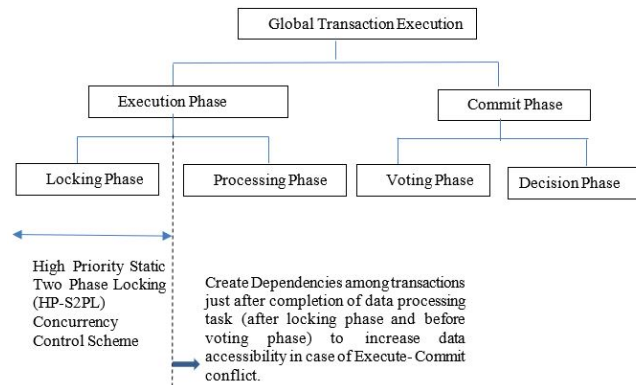
Health Factor (HF) parameter is used to predict whether the requested lending is going to be fruitful or not. The HF is computed for each lender to make the lending decision intelligently — only when doing so benefits the transactions that are going to cooperate through the idea of lending-borrowing activity. Mathematically, the HF can be defined as (Haritsa, Ramamritham, & Gupta, 2000).

$$HealthFactor(HF) = \frac{T_{left}}{Min_{time}}$$

Here, T_{left} is the time left until the executing DRTT's deadline, and Min_{time} is the minimum time necessary for commit processing.

In previous protocols, the HF is computed at the time when the coordinator of a DRTT is about to send the PREPARE message to its participating cohorts. In the proposed EDRC protocol, however, the health factor is computed at an earlier time when the cohort is ready to send a WORKSTARTED message to its coordinator. Thereby, it opens the way to start the lending process earlier. The MinHF is a threshold HF value used to decide whether the data held by committing/processing DRTT can be lent to executing (locking) cohort or not. The variable MinHF is the key factor influencing the performance of the EDRC protocol. Therefore, it is always required to have an appropriate value of MinHF for better performance. For $MinHF = 0$, the T_{left} is required to be 0. This means, executing DRTT is ultimately going to miss its deadline, and therefore lending will be of no use. For $MinHF = \infty$, the borrowing feature is disabled; thus, the EDRC protocol will behave like the classic 2PC. When $MinHF = 1$, the fruitful lending is possible. In our experiments, the MinHF is assumed to be 1.2 (Shanker, Misra, & Sarje, 2006) since it is the right value for efficiently controlling the optimistic data access. It is the same as

Figure 4. Level of Data Accessibility in EDRC Protocol



the value considered by all the previous protocols (Qin & Liu, 2003) (Qin, Liu, & Yang, 2003) (QIN & LIU, 2002). The success percentage of the lending-borrowing event is found to be much better than for $\text{MinHF} = 1.2$ as compared to that of $\text{MinHF} = 1$.

Figure 4 represents the level of data accessibility with the EDRC protocol. As it is clearly interpreted from the above figure, the dependency among the set of transactions may occur even when they all are in the execution phase; of course, they are in different sub-phases of execution phases. Such action leads to increased data accessibility which is crucial in today's real-world transactional data processing.

PERFORMANCE EVALUATION

For the experimental assessment of the EDRC protocol, there is no readymade simulation tool available (Ulusoy, 1995) (Taina & Son, 1999). The event-driven simulation is often used in most of the average and substantially large size database applications consisting of transactions executing concurrently and competing for the shared resource(s). Therefore, a discrete event-driven simulation is developed using C language to access the performance of the proposed protocol. The performance of the EDRC protocol is measured using the metric named "Miss Percentage". It is the percentage of transactions that are aborted/killed as a result of their deadline miss. Being an extension of the SWIFT protocol, the model assumptions, simulations parameter and other parameter detail can be obtained by referring (Shanker, Misra, & Sarje, 2006) since analogues environment is created for experimentation and comparison purposes. For this reason, the EDRC protocol's performance is compared with PROMPT, 2SC and SWIFT Protocol. To investigate the performance of the EDRC, when applied to a DRTDBS, a wide range of global as well as local transactions are generated.

Figures 5-7 present the transaction miss percentage in the disk resident databases under resource and data contention at communication delay of either 10 milliseconds or 1 millisecond with different transaction arrival rates.

Figure 5. Transaction Arrival Rate vs Transaction Miss Percentage: 1 millisecond comm. delay and normal load

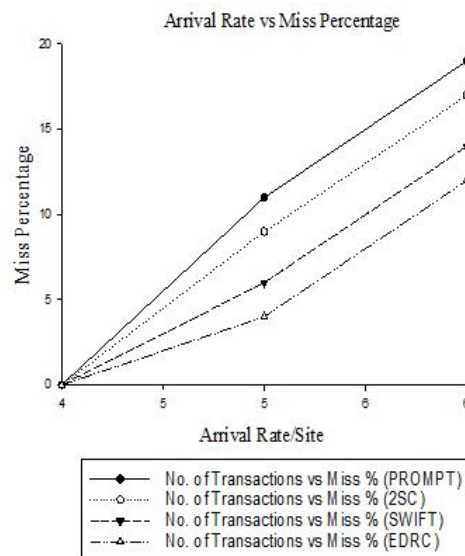


Figure 6. Transaction Arrival Rate vs Transaction Miss Percentage: 1 millisecond comm. delay and heavy load

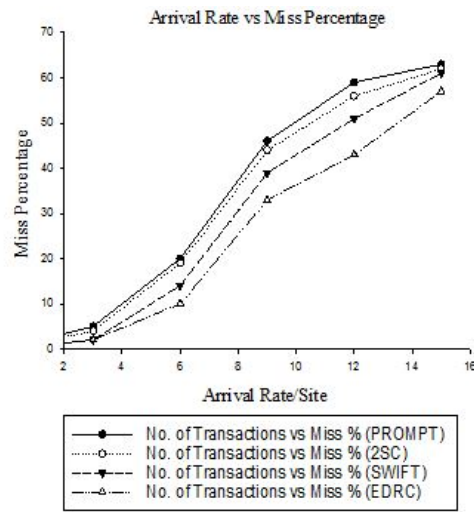
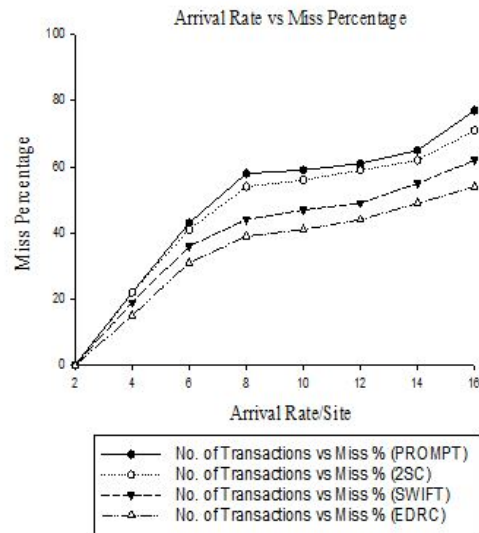


Figure 7. Transaction Arrival Rate vs Transaction Miss Percentage: 10 millisecond comm. delay and normal & heavy load



It is seen that the proposed EDRC protocol performs better than PROMPT, 2SC and SWIFT protocol. The main reasons for the improvement in performance are refined classification and handling of data conflicts among concurrently executing transactions, and thus increased data accessibility in the EDRC protocol.

FUTURE RESEARCH TRENDS

The point-to-point insight about future research opportunities from the DRTDBS perspective is provided below (Pandey & Shanker, 2020).

1. The data conflicts (locking-locking, locking-processing & locking-committing conflict) and possible priority inversion because of them are one of the main reasons for degraded system performance. Therefore, existing transaction scheduling algorithms need a fresh look from the perspective of efficiently resolving data conflicts. The priority inheritance approach has a potential to become a promising approach in dealing with priority inversion while scheduling DRTTs as it can be utilized to improve existing concurrency control protocols and commit protocols (Pandey & Shanker, 2017b) (Pandey & Shanker, 2018a) (Pandey & Shanker, 2018b) (Pandey & Shanker, 2018c).
2. Only a few research works have been done in the context of the handling of non-database operations; though, research in this direction may be useful to improve underlying system performance. The Non-Database Operations Aware Priority Ceiling Protocol (N-DOAP) is a recent effort in this direction (Singh, Pandey, & Shanker, 2019). In the N-DOAP protocol, the database operations are given preference over non-database operations. Consideration of the aspect of non-database operations resulted in the improvement of performance.
3. In RTDBS with one or more processor, there is a need of an effective and workable approach to reduce both transaction miss percentage and power consumption (Bambagini, Marinoni, Aydin, & Buttazzo, 2016) (Baruah, Bertogna, & Buttazzo, 2015) (Guo, Bhuiyan, Saifullah, Guan, & Xiong, 2017). One way to do so is by merging the similar real-time queries for the decrement of repeated data processing and accessing while doing power management dynamically (Kang K., Reducing deadline misses and power consumption in real-time databases, 2016) (Kang & Chung, 2017). It is still not clear how a one deadline can be assigned to a query in real time environment that in fact combines many real-time queries with the outset of different deadlines into one (Madden, Franklin, Hellerstein, & Hong, 2005) avoiding priority inversion (Kang K., 2018). Related work in this area is relatively scarce in spite of being a very important one.
4. It will be interesting to widen the discussion of priority inversion problem of mobile DRTDBS (Lam, Kuo, Tsang, & Law, 2000) (Lei, Zhao, Chen, & Yuan, 2009) (Swaroop, Gupta, & Shanker, 2011), replicated DRTDBS (Xiong, Ramamritham, Haritsa, & Stankovic, 2002) (Wei, Aslinger, Son, & Stankovic, 2004, August) (Srivastava, Shankar, & Tiwari, 2012), nested (Abdouli, Sadeg, Amanton, & Alimi, 2005, September) (Moon & Lee, 2011), active DRTDBS (Lam, Law, & Lee, 2000), and mobile ad-hoc network (MANET) databases (Xing & Gruenwald, 2011, August) (Xing & Gruenwald, 2013). More promising research efforts are required to develop advanced value based real-time commercial products.
5. Considering the wide-spreading impact of database system based applications in recent days, researchers also started working to improve the performance of conventional databases systems by improving classic 2PC and 2PL (Gupta & Sadoghi, March, 2018) (Gupta & Sadoghi, 2019) (Pandey, Pandey, & Shanker, 2019) (Harding, Aken, Pavlo, & Stonebraker, 2016).

CONCLUSION

The proposed EDRC protocol optimistically utilized the Lender-Borrower approach. It recategorized the data conflict as a locking-locking conflict, locking-processing conflict, and locking-committing conflict. The novel feature of this protocol is that it facilitates to start the lending process just after the completion of the data processing task of the cohort. For guaranteeing the feature of early lending, the health factor of a cohort is computed at a time when it sends WORKSTARTED message to its coordinator. This overlaps the execution and commit phase, and thereby reduces the overall completion time of global transaction. As a result, it reduced the transaction deadline miss percentage.

Future research in the direction of advancement of existing state-of-the-art lender-borrower approach based transaction scheduling protocols is required to come up with advanced user oriented commercial products/services.

ACKNOWLEDGMENT

The support of the Council of Scientific and Industrial Research, New Delhi, India, Grant no 1061461137, is acknowledged for this research effort.

REFERENCES

- Aakash, A., Gaurav, A., Gupta, A., & Kumar, B. (2015). Extended SWIFT: A Real Time Commit Protocol. *SCIT Journal*, 15.
- Abdouli, M., Sadeg, B., Amanton, L., & Alimi, A. (2005, September). A system supporting nested transactions in DRTDBSs. In *International Conference on High Performance Computing and Communications*. Springer. 10.1007/11557654_99
- Bambagini, M., Marinoni, M., Aydin, H., & Buttazzo, G. (2016). Energy-aware scheduling for real-time systems: A survey. *ACM Transactions on Embedded Computing Systems*, 15(1), 1–34. doi:10.1145/2808231
- Baruah, S., Bertogna, M., & Buttazzo, G. (2015). *Multiprocessor scheduling for real-time systems*. Springer. doi:10.1007/978-3-319-08696-5
- Guo, Z., Bhuiyan, A., Saifullah, A., Guan, N., & Xiong, H. (2017). Energy-efficient multi-core scheduling for real-time DAG tasks. *LIPICs-Leibniz International Proceedings in Informatics*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 76.
- Gupta, R., Haritsa, J., & Ramamritham, K. (1997). Revisiting commit processing in distributed database systems. *SIGMOD Record*, 26(2), 486–497. doi:10.1145/253262.253366
- Gupta, R., Haritsa, J., Ramamritham, K., & Seshadri, S. (1996). Commit Processing in Distributed Real-Time Database Systems. *Real-Time Systems Symposium*, 220-229. 10.1109/REAL.1996.563719
- Gupta, S., & Sadoghi, M. (March, 2018). EasyCommit: A Non-blocking Two-phase Commit Protocol. *International Conference on Extending Database Technology (EDBT)*, 157-168.

- Gupta, S., & Sadoghi, M. (2019). Efficient and non-blocking agreement protocols. *Distributed and Parallel Databases*, 1–47.
- Harding, R., Aken, D. V., Pavlo, A., & Stonebraker, M. (2016). An Evaluation of Distributed Concurrency Control. *Vldb*, 10(05), 553–564.
- Haritsa, J. R., Ramamritham, K., & Gupta, R. (2000). The PROMPT real-time commit protocol. *IEEE Transactions on Parallel and Distributed Systems*, 11(02), 160–181. doi:10.1109/71.841752
- Kang, K. (2016). Reducing deadline misses and power consumption in real-time databases. *IEEE Real-Time Systems Symposium (RTSS)*, 257–268. 10.1109/RTSS.2016.033
- Kang, K. (2018). Enhancing timeliness and saving power in real-time databases. *Real-Time Systems*, 54(2), 484–513. doi:10.1007/11241-018-9302-2
- Kang, W., & Chung, J. (2017). Energy-efficient response time management for embedded databases. *Real-Time Systems*, 53(2), 228–253. doi:10.1007/11241-016-9264-1
- Lam, K., Kuo, T., Tsang, W., & Law, G. (2000). Concurrency control in mobile distributed real-time database systems. *Information Systems*, 25(4), 261–286. doi:10.1016/S0306-4379(00)00018-1
- Lam, K., Law, G., & Lee, V. (2000). Priority and deadline assignment to triggered transactions in distributed real-time active databases. *Journal of Systems and Software*, 51(1), 49–60. doi:10.1016/S0164-1212(99)00109-0
- Lei, X., Zhao, Y., Chen, S., & Yuan, X. (2009). Concurrency control in mobile distributed real-time database systems. *Journal of Parallel and Distributed Computing*, 69(10), 866–876. doi:10.1016/j.jpdc.2009.06.008
- Madden, S., Franklin, M., Hellerstein, J., & Hong, W. (2005). TinyDB: An acquisitional query processing system for sensor networks. [TODS]. *ACM Transactions on Database Systems*, 30(1), 122–173. doi:10.1145/1061318.1061322
- Moon, S., & Lee, S. (2011). A Reliable Nested Transaction Model with Extension of Real-Time Characteristics. In *Reliable and Autonomous Computational Science* (pp. 123–142). Springer. doi:10.1007/978-3-0348-0031-0_7
- Pandey, A. K., Pandey, S., & Shanker, U. (2019). LIFT- A New Linear Two-Phase Commit Protocol. *Proceedings of 25th annual International Conference on Advanced Computing and Communications (ADCOM 2019) at IIT Bangalore*.
- Pandey, S., & Shanker, U. (2016). Transaction Execution in Distributed Real-Time Database Systems. *Proceedings of the International Conference on Innovations in information Embedded and Communication Systems*, 96–100.
- Pandey, S., & Shanker, U. (2017a). IDRC: A Distributed Real- Time Commit Protocol. *Proceedings of the 6th International Conference on Smart Computing and Communications, (ICSCC) Kurukshetra, India*.
- Pandey, S., & Shanker, U. (2017b). On Using Priority Inheritance Based Distributed Static Two Phase Locking Protocol. *Proceedings of the International Conference on Data and Information System (IC-DIS)*, 179–188.

- Pandey, S., & Shanker, U. (2018a). Priority Inversion in DRTDBS: Challenges and Resolutions. *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD '18)*, 305-309. 10.1145/3152494.3167976
- Pandey, S., & Shanker, U. (2018b). CART: A Real-Time Concurrency Control Protocol. In *22nd International Database Engineering & Applications Symposium (IDEAS 2018)*. ACM.
- Pandey, S., & Shanker, U. (2018c). A One Phase Priority Inheritance Commit Protocol. *Proceedings of the 14th International Conference on Distributed Computing and Information Technology (ICDCIT) Bhubaneswar, India, January 11-13 2018*. 10.1007/978-3-319-72344-0_24
- Pandey, S., & Shanker, U. (2020). Transaction Scheduling Protocols For Controlling Priority Inversion: A Review. *Computer Science Review*, 35, 100215. doi:10.1016/j.cosrev.2019.100215
- QIN, B., & LIU, Y. S. (2002). Distributed real-time transaction commit processing. *Journal of Software*, 13(8), 1395–1401.
- Qin, B., & Liu, Y. (2003). High performance distributed real-time commit protocol. *Journal of Systems and Software*, 68(02), 145–152. doi:10.1016/S0164-1212(02)00145-0
- Qin, B., Liu, Y., & Yang, J. (2003). A commit strategy for distributed real-time transaction. *Journal of Computer Science and Technology*, 18(5), 626–631. doi:10.1007/BF02947122
- Shanker, U. (2008). *Some Performance Issues in Distributed Real Time Database Systems* (PhD Thesis). Indian Institute of Technology Roorkee.
- Shanker, U., Agarwal, N., Tiwari, S., Goel, P., & Srivastava, P. (2010). ACTIVE-a real time commit protocol. *Wireless Sensor Network*, 2(3).
- Shanker, U., Misra, M., & Sarje, A. (2001). Hard Real Time Distributed Database Systems: Future Directions. *Proceedings of All India Seminar on Recent Trends in Computer Communication Networks, Dept. of ECE, IIT Roorkee, India*, 172-177.
- Shanker, U., Misra, M., & Sarje, A. K. (2006). SWIFT - A new real time commit protocol. *Distributed and Parallel Databases*, 20(01), 29–56. doi:10.1007/10619-006-8594-8
- Shanker, U., Misra, M., & Sarje, A. K. (2008). Distributed real time database systems: Background and literature review. *International Journal of Distributed and Parallel Databases, Springer Verlag*, 23(02), 127–149. doi:10.1007/10619-008-7024-5
- Shanker, U., Misra, M., & Sarje, A. K. (2009). SWIFT: A Distributed Real Time Commit Protocol. In *Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends* (pp. 737-743). IGI Global. doi:10.4018/978-1-60566-242-8.ch078
- Shanker, U., Vidyareddi, B., & Shukla, A. (2012). PERDURABLE: A real time commit protocol. *Recent Trends in Information Reuse and Integration*, 1-17.
- Singh, R. K., Pandey, S., & Shanker, U. (2019). A Non-Database Operations Aware Priority Ceiling Protocol for Hard Real-Time Database Systems. *Proceedings of 10th International Conference on Computing Communication and Networking Technologies*.

Srivastava, A., Shankar, U., & Tiwari, S. K. (2012). Transaction management in homogenous distributed real-time replicated database systems. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6).

Swaroop, V., Gupta, G. K., & Shanker, U. (2011). Issues in mobile distributed real time databases: Performance and review. *International Journal of Engineering Science and Technology*, 3(4).

Taina, J., & Son, S. H. (1999). Towards a general real-time database simulator software library. *IFAC Proceedings*, 32(1), 75-80.

Ulusoy, O. (1995). A study of two transaction-processing architectures for distributed real-time database systems. *Journal of Systems and Software*, 31(02), 97–108. doi:10.1016/0164-1212(94)00090-A

Wei, Y., Aslinger, A., Son, S., & Stankovic, J. (2004, August). ORDER: A dynamic replication algorithm for periodic transactions in distributed real-time databases. *10th International Conference on Real-Time and Embedded Computing Systems and Applications (RTCS 2004)*.

Xing, Z., & Gruenwald, L. (2011, August). An energy-efficient concurrency control algorithm for mobile ad-hoc network databases. In *International Conference on Database and Expert Systems Applications*. Springer. 10.1007/978-3-642-23088-2_37

Xing, Z., & Gruenwald, L. (2013). Managing concurrent execution of transactions in mobile ad-hoc network database systems: An energy-efficient approach. *Distributed and Parallel Databases*, 31(2), 183–230. doi:10.1007/10619-012-7114-2

Xiong, M., Ramamritham, K., Haritsa, J. R., & Stankovic, J. A. (2002). MIRROR: A state-conscious concurrency control protocol for replicated real-time databases. *Information Systems*, 27(04), 277–297. doi:10.1016/S0306-4379(01)00053-9

ADDITIONAL READING

Bernstein, P., & Newcomer, E. (2009). *Principles of transaction processing*. Morgan Kaufmann.

Elmasri, R. (2008). *Fundamentals of database systems*. Pearson Education India.

Gray, J. (1981, September). The transaction concept: Virtues and limitations. In VLDB, 81, 144-154.

Gray, J., & Reuter, A. (1992). *Transaction processing: concepts and techniques*. Elsevier.

Gupta, S., & Sadoghi, M. (2018). Blockchain Transaction Processing.

Kim, Y. (1995). *Predictability and consistency in real-time transaction processing*. Doctoral dissertation, University of Virginia.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>

Ramamritham, K., Son, S. H., & Dipippo, L. C. (2004). Real-time databases and data services. *Real-Time Systems*, 28(2-3), 179–215. doi:10.1023/B:TIME.0000045317.37980.a5

Sadoghi, M., & Blanas, S. (2019). Transaction Processing on Modern Hardware. *Synthesis Lectures on Data Management*, 14(2), 1–138. doi:10.2200/S00896ED1V01Y201901DTM058

Shanker, U., Misra, M., & Sarje, A. (2006). Some performance issues in distributed real-time database systems. *Proc. VLDB Ph.D. Work, Conv. Exhib. Cent. (COEX), Seoul, Korea*.

Yu, P. S., Wu, K., Lin, K., & Son, S. H. (1994). On Real-Time Databases: Concurrency Control and Scheduling. *Proceedings of the IEEE*, 82(01), 140–157. doi:10.1109/5.259432

KEY TERMS AND DEFINITIONS

Distributed Real-Time Database System (DRTDBS): The DRTDBS consists of multiple data sites that are connected through an underlying computer network specifically used to run applications based on distributed real-time transactions.

Firm Distributed Real-Time Transaction (Firm DRTT): The firm DRTT is killed in case it misses its deadline because its outcome has no value after the deadline miss.

Hard Distributed Real-Time Transaction (Hard DRTT): The hard DRTT must be finished before its deadline; otherwise, it may lead to a potentially catastrophic consequence.

Locking-Committing Conflict: The locking-committing conflict is handled in the same way as execute-commit conflict.

Locking-Locking Conflict: The locking-locking conflict is handled simply by aborting the DRTT with low priority in case of priority inversion.

Locking-Processing Conflict: The locking-processing conflict is handled based on the status of the cohort i.e. whether it has sent WORKSTARTED message or not, and lending may be allowed only when the cohort has already sent WORKSTARTED message.

Real-Time Commit Protocols: Ensure that either all the effects of the DRTT persist or none of them irrespective of any such failure.

Soft Distributed Real-Time Transaction (Soft DRTT): The soft DRTT is not killed/aborted in case of its deadline miss because the result has some value (obviously degrading) even after the deadline miss.

Application of Data Mining in Smart Grid Technology

6

Lipi Chhaya*Researcher, University of Petroleum and Energy Studies, India***Paawan Sharma***Pandit Deendayal Petroleum University, India***Adesh Kumar***University of Petroleum and Energy Studies, India***Govind Bhagwatikar***Sany Wind Energy India Private Limited, India*

INTRODUCTION

Smart grid technology is a radical and creative approach for improvisation in existing passive power grid. Assimilation of power and communication infrastructures is inevitable for the design and deployment of an imminent smart grid technology. The grid is called ‘Smart’ as it is accomplished with intelligent devices for decision making. Some of the substantial features of smart grid technology are bidirectional communication, renewable integration, active consumer participation, distribution automation, advanced metering infrastructure, supervisory control and data acquisition and complete management of entire power grid through wireless communication protocols and technologies. Smart grid communication infrastructure is a gigantic architecture with massive amount of data communication between various hierarchical network layers (Farooq & Jung, 2014). Full duplex data communication is required for monitoring, control, security, management and fault diagnosis of entire power grid (Mahmood et al., 2015). Management of complex, hierarchical and heterogeneous smart grid communication infrastructure necessitates data collection, storage, processing, analysis, retrieval and communication for self-healing and complete automation. Collection, interpretation and analysis of massive quantity of data generated in smart grid operation are critical, inevitable and complex tasks (Gungor et al., 2010). Data mining techniques can be an effective solution for smart grid operation and management. Data mining is essential for converting information into knowledge (Wijaya, T., 2013). Knowledge discovery and data mining are interdisciplinary tasks in the context of smart grid network. Database management and data mining are the two significant aspects of a database system. Database management deals with processing and storage of data and data mining deals with discovery and abstraction of knowledge for the purpose of decision making. Data mining is a computational process for data scrutiny and analysis. Data analysis is unavoidable for unambiguous knowledge discovery as well as decision making practices. Data mining is necessary for analysis of various statistics associated with power generation, distribution automation, data communications, billing, consumer participation, and fault diagnosis in smart power grid. In Smart grid, accumulation of real time data is inevitable. Data mining techniques are required when data is continuously collected on real time basis (Atzmueller et al., 2013). The pattern of gigantic sets of data

DOI: 10.4018/978-1-7998-3479-3.ch056

can be effectively extracted and scrutinized with the use of different data mining techniques for energy efficiency, reliability and real time decision making. This chapter is expected to serve as a comprehensive analysis and review of application of data mining techniques in generation, transmission, distribution and utilization of energy in smart grid infrastructure.

BACKGROUND

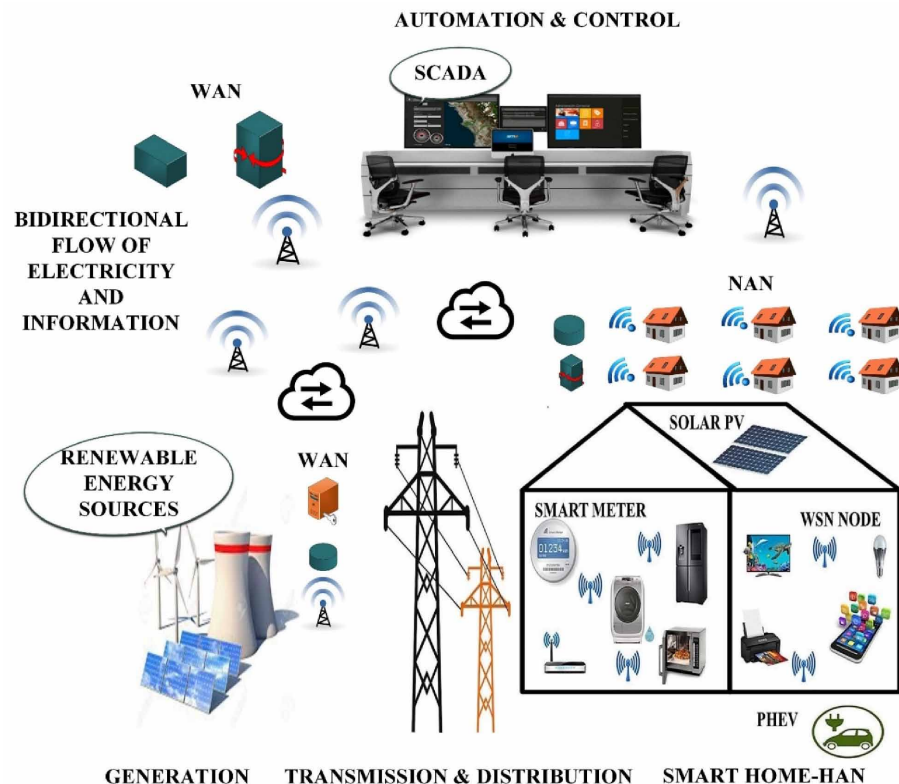
An electrical power grid carries the electricity from generation unit to the consumption unit. An existing power grid is being transformed through the design, development and deployment of smart grid technology. An existing grid is passive in terms of real time monitoring, control, consumer participation, fault diagnosis, disaster recovery, and real time communication of energy related statistics. There are enormous challenges such as theft and distribution losses, scarcity of primary energy sources, aging and underperformance of coal based power plants etc. An existing grid is incapable of addressing these issue in effective manner due to lack of data communication between various components of energy generation, transmission, distribution and utilization. Smart grid can be defined in diverse ways as per its functional, technological or constructive aspects. (United States Department of Energy, 2009) defines the Smart Grid as: 'A smart grid uses digital technology to improve reliability, security, and efficiency (both economic and energy) of the electric system from large generation, through the delivery systems to electricity consumers and a growing number of distributed-generation and storage resources.' Smart Grid includes various communications protocols and hierarchical network layers. (Farooq & Jung, 2014) have discussed various choices available for realization of smart grid communication networks. The hierarchical design of Smart power grid comprises of home, neighborhood and wide area networks. These hierarchical networks exchange information for monitoring and control of smart grid operations. Mahmood et al. have discussed and reviewed various wireless communication protocols for smart grid applications. Massive quantity of data communication is required for the realization of smart grid technology. An unambiguous knowledge discovery along with data analysis is vital for reliable and proficient management of smart grid. The prime concern is an extraction, analysis, interpretation and management of exploded data from various measurement devices and sensors in smart grid. Smart grid is a multifaceted technology. Moreover, the nature of data in smart grid technology is dynamic and distributed (Saputro et al., 2012). The traditional data mining processes are focused, centralized, static, bounded and non-real time. It does not have energy or speed constraints. Smart grid technology incorporates wireless sensor networks and Internet of Things which are two diverse areas in terms of data mining and knowledge detection processes. The traditional data mining techniques cannot be directly applied to wireless sensor networks due to their restricted computational, power and storage capabilities. Internet of things is meant for dynamic, swift and near real time data which necessitates novel data mining approach. This chapter illustrates conceptual description of smart grid technology and application of data mining techniques in generation, transmission, distribution and utilization of electrical energy.

CONCEPT OF SMART GRID TECHNOLOGY

An existing electrical grid is being transformed through deployment of smart grid technology. The main difference between an existing power grid and smart grid is in terms of full duplex communication and real time management of energy statistics. Smart grid is an assimilation of power as well as information

and communication infrastructures (Erol-Kantarci et al., 2011). The key features of smart grid technology includes real time monitoring, management and control of energy parameters, high penetration of renewable energy resources, advanced metering infrastructure, supervisory control and data acquisition, fault diagnosis, disaster recovery, self-healing, active consumer participation, demand side management, distribution automation and reduction of greenhouse gas emissions (Chen et al., 2006). Smart grid comprises of various hierarchical communication network layers such as home area network, neighborhood area network and wide area network. Home area network is meant for consumer premises. The range of this network is few hundred meters and it facilitates the communication between utility and consumers (Sun et al., 2009). Home area network consists of communication transmitters, receivers, intelligent electronic devices, consumer electronic devices, energy meters and renewable energy sources. Neighborhood area network consist of several home area networks. The data collected by smart meters in home area network is forwarded to wide area network through neighborhood area network. The area of neighborhood area network is few kilometers. Wireless local area network, cellular communication standards and power line communication technologies can be used for implementation of neighborhood area network (Patel et al., 2011). Wide area network connects several neighborhood area networks. Wide area requires very high bandwidth for its operation and management. Wide area network is appropriate for supervisory control and data acquisition systems for monitoring, data acquisition, control, and management of power grid. Internet of things technology is inevitable at each stage of hierarchical network for connectivity and integration of various equipment and meters. Various cellular communication standards and power line communication technologies are suitable choices for realization of wide area network (Chhaya et al.,

Figure 1. Smart grid communication infrastructure



2017). Hierarchical architecture ensures reliable communication of information at generation, transmission, distribution and control of smart grid but still there are enormous challenges to be overcome for realization of full-fledged smart grid communication infrastructure. Realization of hierarchical network necessitates massive data communication and analysis for its operational security and reliability (Xu et al., 2014). Data mining tasks and techniques are inevitable for functional efficiency of Smart Grid networks. Figure 1 shows the architecture of smart grid communication network.

PROCESS OF DATA MINING

Data mining is expedient for optimization of Smart Grid network. As functional and operational efficiency of smart grid is based on decision making, different data mining techniques are suitable for generation, transmission, distribution and consumption stages of smart power system. The mining of information from large volume of data is done through the methods from database systems, artificial intelligence, machine learning and statistics. Realization of Smart grid is reliant on integration of information and communication technology with power grid as fundamental concepts and operations of mechanical and electrical apparatus will remain unchanged. The innovative approach is to integrate these devices with smart sensors and actuators for extraction and communication of information. This information is useful for monitoring and control purpose. The patterns of gigantic sets of data can be effectively extracted and analyzed with the use of data mining techniques which results into energy efficiency, reliability and real time decision making. The use of data mining technology is to discover reliable, valid and understandable information. Figure 2 shows the process of data mining.

Data mining can be defined as a computational process to extract and scrutinise the information from massive amount of data sets by using techniques such as artificial intelligence, machine learning, database systems, and statistics (Jiawei et al., 2000). It is a collaborative and iterative process of knowledge discovery. Discovery of patterns from huge sets of data generally includes following tasks.

- **Clustering:** The process of clustering identifies similar objects in the data and clusters the same. It is a grouping of identical objects in huge sets of data.
- **Anomaly Detection:** This process identifies deviation or unusual data.
- **Classification:** It is a process of generalization of known structure for application of novel data.
- **Summarization:** It deals with report generation and visualization by compact depiction of data sets.
- **Regression:** It finds the function for modelling of data with minimum errors.
- **Association rule learning:** This process finds reliance or connection between different variables.
- **Frequent Pattern mining:** It is a process to find the motifs or patterns which occur frequently during pattern extracting process.

Data mining process can be rarely performed with the data sets in their original forms. So, data cleaning, aggregation, normalization, identification of noisy values and compression can be performed before data mining process. Various data mining methods are used in combination or independently to perform various data mining tasks.

Figure 2. Process of data mining

APPLICATION OF DATA MINING IN SMART GRID

Data mining tasks and methods are required to identify suitable input as well as output variables and to extract the features of these variables. Data mining methods can be used to identify power system conditions such as emergency, normal, restorative or alert (Mahmood et al., 2013).

Applications of Data Mining in Generation

As the nature of various generation sources is quite diverse, data mining has extensive applications in generation. Association rule mining can be used for fault identification and troubleshooting of coal based power plant. Statistical analysis can be used to analyse the performance and operational characteristics of renewable energy sources due to their unpredictable features (Cai, T., 2016). Photovoltaic generation

system behaviour and failure analysis can be performed using data mining techniques (Serrano-Luján et al, 2015). Various regression tasks can be performed for load profile forecasting. In fossil fuel based power plant, supply and demand can be synchronised by load forecasting which may result into saving of revenue. This method can also be applied to renewable energy resources. Load forecasting can be performed for very short term (for hours), medium term (for few weeks), or long term (for years). Load forecasting can be performed using multiple regression, artificial neural network, or time series.

Applications of Data Mining in Transmission and Distribution

Wireless sensor is a vital component of smart grid network. Wireless sensor networks are used for collection and communication of data among smart grid communication networks. Wireless sensor network comprises of sensor, storage unit, communication transceivers, power supply unit, location finding system and actuator. Wireless sensor networks are meant for real time monitoring and control of energy statistics in smart grid network (Chhaya et al., 2017).

Estimation of power system is performed for evaluation of metered as well as unmetered connections and quantities of whole power system. It is useful for power failure prevention and grid stability. Association rule mining and fuzzy logic can be used for power system estimation. Transformer insulation oil is stable at high temperature and retains excellent insulation properties. Dissolved Gas Analysis is used to estimate the amount of dissolved gases in transformer insulating oil. It is useful for fault diagnosis and maintenance of transformer. Artificial intelligence can be applied to perform dissolved gas analysis. Integration of renewable energy resources is the most prominent feature of Smart Grid in the context of reduction of GHG emissions and energy efficiency. Micro grid facilitates the high penetration of renewable energy resources. It can operate in two modes namely grid connected and island mode. In island mode, renewable energy resources self-reliantly provide power to consumers. Islanding can also be used when the grid power is not available or there is a fault in the network. Island detection is also important for grid stability when many distributed energy sources are connected to power grid. Island detection can be performed using fuzzy membership function using rule based approach. Fault identification can be performed using artificial neural networks and regression processes. Data mining techniques are useful for secured metering system (Faisal et al., 2012).

Power quality is a crucial aspect of Smart Grid system. Power quality can be evaluated in terms of various electrical parameters such as surges, spikes, harmonics, droops, sags, and instabilities (Hongke & Linhai, 2010). Online analytical processing can be performed for power quality evaluation. Contingency analysis is performed for identification of potential failures in the power system and their effect on the system. This analysis can also be used to evaluate the tolerance of power system against various fault conditions. Electricity price forecasting is an important parameter for market operation. It is imperative for energy distributors as well as consumers. Regression techniques and artificial neural networks can be used for forecasting of energy prices. Active consumer participation is a key characteristic of Smart grid network. Consumer can monitor and manage their energy utilization through real time communication of information. Dynamic billing cycle can be used through load forecasting. Short duration load forecasting is possible using data mining techniques. This forecasting majorly impacts future energy consumption pattern and dynamic pricing. Neural network can be used for electricity load and price forecasting (Zahid et al., 2019).

Applications of Data Mining in Utilization

Consumer is of prime importance in Smart Grid network. Smart Grid technology facilitates an active participation of consumers. Consumers can monitor and control their energy usage as they will be continuously updated with various metering related information. They can optimize their charge of energy usage by getting information regarding time of the day billing information. Advanced metering infrastructure is also an inevitable part of Smart Grid. It can be defined as a full duplex network between utilities and customer premises. Security of advanced metering infrastructure network is the most critical aspect of smart grid network. National Institute of Standards and Technology has proposed guidelines for smart grid cyber security issues. (Faisal et al., 2012) have proposed stream data mining techniques and algorithms for intrusion detection system for advanced metering infrastructure network. Classification methods of data mining can also be used for advanced metering infrastructure network security and attack and intrusion resilience. Load profiling and forecasting are very important parameters for energy efficiency and reliability. Smart micro grid enhances the penetration of renewable energy resources in power system. Data mining techniques are apt solution for analysis of uncertain behaviour of renewable energy resources (Cai, T., 2016). In case of Micro grids, where user or a community can select various parameters of Grid, renewable operations and time of usage, load forecasting is a crucial requirement. It is also important for Building Management System. The comparative study of seven data mining techniques for load forecasting in Building Management System is explored in. Support vector machine, autoregressive method, neural networks polynomial model etc. can be used for building load forecasting. Electricity price forecasting for consumer applications can be performed using regression techniques. Different hybrid data mining methods can also be developed for various applications in Smart Grid.

Decision tree is used to explore the hidden relation relationship between input and output. It is used to predict and assess the operating conditions of power system. It can also be used for control, protection, fault diagnosis and load forecasting of power distribution system (Liu et al., 2013).

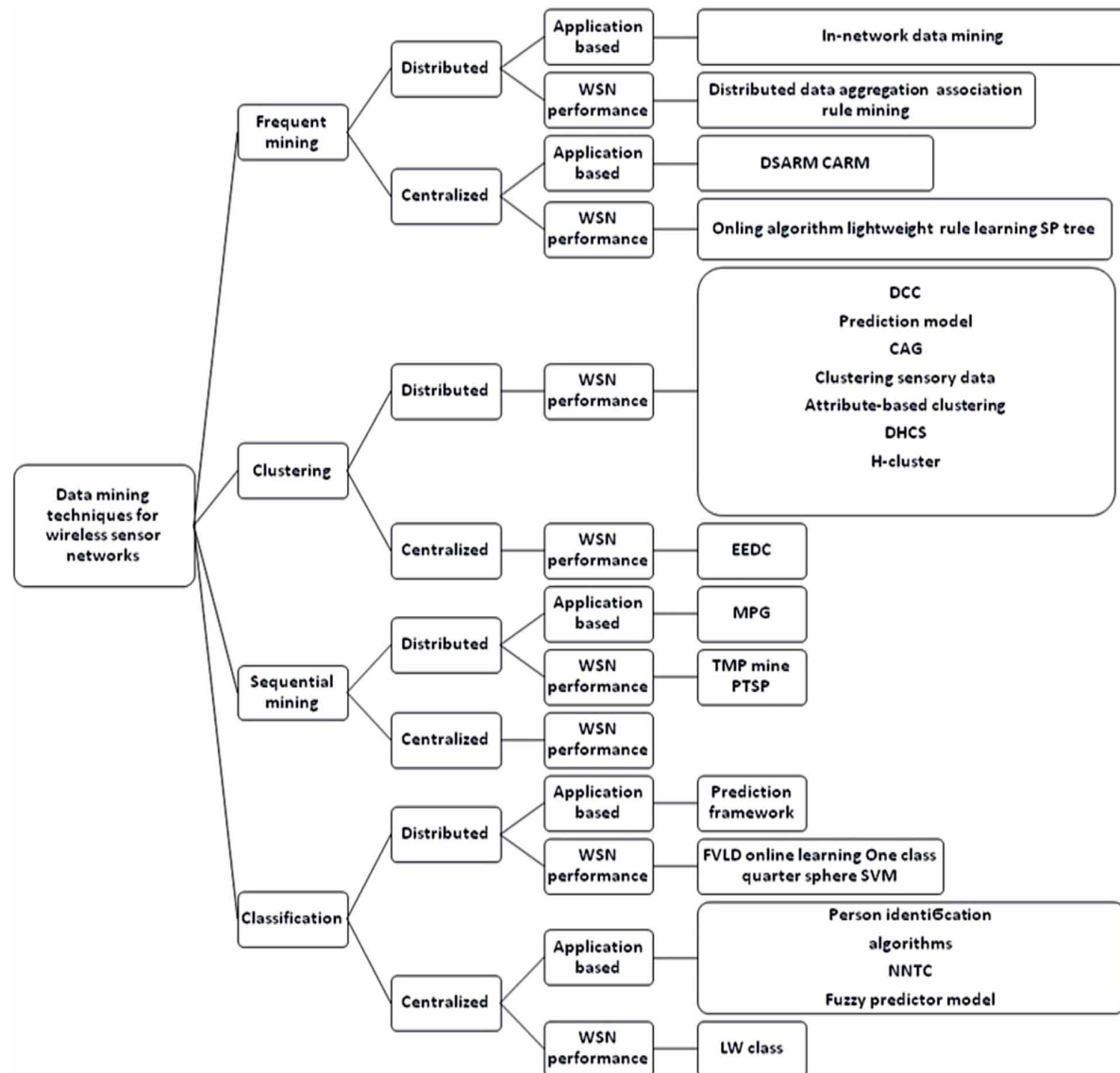
Data Mining Techniques for Wireless Sensor Networks

Wireless sensor network is a vital part of smart grid communication network as it communicates real time energy statistics to different smart grid hierarchical network layers. Wireless sensor network comprises of communication transceivers, storage unit, power supply unit, processing unit, sensors, and actuator. Wireless sensor network is an inevitable part of different smart grid network layers. Furthermore wireless sensor network is a vulnerable entity due to its limited storage, processing and computational capabilities. The traditional data mining techniques are used for static, non-real time, centralized and bounded networks with high memory, computing and power capabilities. Moreover, these networks do not require very high speed communications. Wireless sensor networks have limited capabilities and require the processing of real time and high speed dynamic data (Mahmood et al., 2013). Figure 3 illustrates the classification of various data mining techniques for wireless sensor network.

Wireless sensor comprises of enormous number of nodes for diverse set of communication standards and networks for smart grid heterogeneous and hierarchical networks. The flow of data is swift, continuous and dynamic. Wireless sensor network is an integral component of Internet of Things which forms a complex and heterogeneous smart grid network.

The purpose of data mining techniques is to extract useful and understandable information from the fast and continuous flow of data. In the centralized approach of data mining, the data is collected at a single central site and then the algorithms are applied for extraction of information from the accumulated

Figure 3. Classification of data mining techniques for wireless sensor network



data. It is like a warehousing and mining of data. This approach is inappropriate as well as challenging due to reasons like gigantic amount of data, bandwidth limitations, storage issues and computational restrictions. The distributed processing of data overcomes the challenges imposed by centralized data mining. Distributed data mining reduces the issues related to storage, power and bandwidth. Smart grid applications are time-critical as real time data is to be communicated to central controller. Distributed data mining techniques prevents gigantic set of raw data to be sent to sink node which is beneficial for constrained wireless sensor networks (Mahmood et al., 2013). Furthermore, there is a trade-off between the performance of wireless sensor network in terms of restricted capabilities and its performance based on specific application. Wireless sensor network performance requires better storage, computational and energy capabilities while application oriented approach requires accuracy, reliability, robustness scal-

ability and fault tolerance of the network. Thus the classification of data mining techniques is carried out on the basis of performance versus application based approach.

CHALLENGES IN DATA MINING OF SMART GRID BIG DATA

Smart grid is a complex hierarchical network with heterogeneous networks and standards. Huge amount of information exchange between diverse intelligent electronic devices and networks is essential for operation and management of smart grid network. Extraction of useful information and analysis of data is a challenging task.

- Handling the gigantic volume of data is the biggest challenge to overcome as knowledge extraction from terabytes, petabytes or zettabytes of data requires proficient algorithms and data mining techniques.
- It is a challenging task to handle heterogeneous data types retrieved from heterogeneous networks, devices, and sources. With an inevitable inclusion of Internet of things in smart grid technology, Internet will be flooded with information communicated between diverse IP based devices and networks.
- The format as well as volume of data is very complex and thus it is challenging to extract an information from gigantic set of heterogeneous and complex data.
- Smart grid comprises of diverse set of networks with heterogeneous standards and technologies. The biggest complexity lies in handling the data from wireless sensor network which is a constrained network in terms of memory, computational proficiency, and power. It is difficult to maintain reliability and accuracy of data from this type of networks.
- The data communication in smart grid hierarchical; networks is very fast as it is near real time data communication. The processing and knowledge discovery from this kind of enormous, rapid and real time data is difficult and complicated task.
- Internet of things is an inevitable component of smart grid network which requires online data mining. Existing data mining and knowledge discovery techniques are offline in nature. The challenge is to handle and process distributed and online data stream.
- Wireless sensor nodes are spreaded over remote and harsh locations. It is challenging to deal with security, heterogeneity, noise, uncertainty, and dynamics of data. Fault finding and correction of incomplete and insufficient data is also challenging.

SOLUTIONS AND RECOMMENDATIONS

The traditional data mining techniques are not suitable for smart grid applications as they are meant for static, non-real time, multipass, bounded and low speed data with no constraints related to network, bandwidth, speed, computations and power. The novel techniques should be developed for offline knowledge discovery and online as well as distributed data processing for Internet of things and wireless sensor network based smart grid network. Data mining and knowledge discovery techniques must consider resource constraints and network topology of wireless sensor nodes. The techniques should be developed for uncomplicated computational algorithms and techniques. Heterogeneous data communication between networks and attribute, spatial and temporal correlations between various sensor nodes

should be consider while developing new data mining and information extraction techniques. Hybrid data mining techniques with single pass algorithms and local models should be developed.

FUTURE RESEARCH DIRECTIONS

Smart grid communication infrastructure is a hierarchical and complex network comprising of heterogeneous communication devices and networks. Data mining and knowledge discovery is the process of information extraction from the gigantic set of raw data. This chapter is expected to serve as a directive for future research endeavors in the field of big data mining techniques. The future research is required for development of data mining algorithms and techniques for wireless sensor networks, Internet of things and heterogeneous networks. Distributed and dynamic approach is required for mining of diverse set of gigantic and real time data. Network dynamics and resource constraints must be consider while developing novel data mining techniques and algorithms. A hybrid approach for secured and fault tolerant information extraction can be considered for future research endeavors.

CONCLUSION

Smart Grid is the most revolutionary technology of present era. An integration of Information and communication technology is inevitable to augment intelligence to smart grid. Extraction of accurate information is possible with the use of data mining methods and tasks. The paper illustrates fundamental components of smart grid communication infrastructure. In this paper, application of data mining in power system generation, transmission, distribution, and consumption are discussed. Furthermore, Internet of things and wireless sensor networks are the vital components of future smart grid technology. The traditional data mining techniques developed for static, non-real time, centralized and bounded networks are not suitable for heterogeneous smart grid communication network. There is a need of rigorous research and development for exploration of data mining techniques for smart grid applications. Smart grid security aspects and fault detection and trouble are the key areas for exploration and future research endeavours. This article is expected to serve as a comprehensive review of potential data mining applications in smart grid research and deployment.

REFERENCES

- Atzmueller, M., Beer, S., & Puppe, F. (2013). Data Mining, Validation, and Collaborative Knowledge Capture. In I. Management Association (Ed.), *Data Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1189-1207). Hershey, PA: IGI Global. doi:10.4018/978-1-4666-2455-9.ch061
- Cai, T. (2016). Application of Data Mining and Analysis Techniques for Renewable Energy Network Design and Optimization. In I. Management Association (Ed.), *Business Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 543-557). Hershey, PA: IGI Global. doi:10.4018/978-1-4666-9562-7.ch028
- Chen, B., Wu, M., Yao, S., & Ni, B. (2006). ZigBee Technology and Its Application on Wireless Meter-reading System. In *IEEE International Conference on Industrial Informatics (INDIN 06)*. IEEE Press. 10.1109/INDIN.2006.275820

Chhaya, L., Sharma, P., Bhagwatikar, G., & Kumar, A. (2017). Communication theories and protocols for smart grid hierarchical network. *Journal of Electrical and Electronics Engineering (Oradea)*, 10(1), 43–48.

Chhaya, L., Sharma, P., Bhagwatikar, G., & Kumar, A. (2017). Wireless Sensor Network Based Smart Grid Communications: Cyber Attacks, Intrusion Detection System and Topology Control. *Electronics (Basel)*, 6(1), 5. doi:10.3390/electronics6010005

Erol-Kantarci, M., & Mouftah, H. (2011). Wireless multimedia sensor and actor networks for the next generation power grid. *Ad Hoc Networks*, 9(4), 542–551. doi:10.1016/j.adhoc.2010.08.005

Faisal, M. A., Aung, Z., Williams, J. R., & Sanchez, A. (2012). Securing advanced metering infrastructure using intrusion detection system with data stream mining. *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI)*, 96–111. 10.1007/978-3-642-30428-6_8

Farooq, H., & Jung, L. (2014). Choices available for implementing Smart Grid communication network. In *Computer and Information Sciences* (Vol. 5). ICCOINS.

Gungor, V., Lu, B., & Hancke, G. (2010). Opportunities and Challenges of Wireless Sensor Networks in Smart Grid. *IEEE Transactions on Industrial Electronics*, 57(10), 3557–3564. doi:10.1109/TIE.2009.2039455

Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hongke, H., & Linhai, Q. (2010). Application and research of multidimensional data analysis in power quality. *Proceedings of the 2010 International Conference on Computer Design and Applications (ICCD)*, 1, 390–393. 10.1109/ICCD.2010.5540836

Liu, Rather, Zakir, Chen, & Bak. (2013). An overview of decision tree applied to power systems. *International Journal of Smart Grid and Clean Energy*, 413–419.

Mahmood, A., Javaid, N., & Razzaq, S. (2015). A review of wireless communications for smart grid. *Renewable & Sustainable Energy Reviews*, 41, 248–260. doi:10.1016/j.rser.2014.08.036

Mahmood, A., Shi, K., Khatoon, S., & Xiao, M. (2013). Data Mining Techniques for Wireless Sensor Networks: A Survey. *International Journal of Distributed Sensor Networks*, 9(7), 406316. doi:10.1155/2013/406316

Patel, A., Aparicio, J., Tas, N., Loiacono, M., & Rosca, J. (2011). *Assessing communications technology options for Smart Grid applications*. IEEE Press. doi:10.1109/SmartGridComm.2011.6102303

Saputro, N., Akkaya, K., & Uludag, S. (2012). A survey of routing protocols for smart grid communications. *Computer Networks*, 56(11), 2742–2771. doi:10.1016/j.comnet.2012.03.027

Serrano-Luján, L., Cadenas, J. M., & Urbina, A. (2015). Classification of Failures in Photovoltaic Systems using Data Mining Techniques. In M. Cascales, J. Lozano, A. Arredondo, & C. Corona (Eds.), *Soft Computing Applications for Renewable Energy and Energy Efficiency* (pp. 300–319). IGI Global. doi:10.4018/978-1-4666-6631-3.ch012

Sun, Z., Zhao, T., & Che, N. (2009). Design of electric power monitoring system based on ZigBee and GPRS. In *1st International Symposium on Computer Network and Multimedia Technology (CNMT 2009)*. IEEE Press. 10.1109/CNMT.2009.5374624

U.S. Department of Energy. (2009). *Smart Grid System Report*. Author.

Wijaya, T. (2013). From Data to Knowledge: Data Mining. In I. Management Association (Ed.), *Data Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1150-1162). Hershey, PA: IGI Global. doi:10.4018/978-1-4666-2455-9.ch059

Xu, Z., Xue, Y., & Wong, K. (2014). Recent Advancements on Smart Grids in China. *Electric Power Components and Systems*, 42(3-4), 251–261. doi:10.1080/15325008.2013.862327

Zahid, M., Ahmed, F., Javaid, N., Abbasi, R., Zainab Kazmi, H., Javaid, A., Bilal, M., Akbar, M., & Ilahi, M. (2019). Electricity Price and Load Forecasting using Enhanced Convolutional Neural Network and Enhanced Support Vector Regression in Smart Grids. *Electronics (Basel)*, 8(2), 122. doi:10.3390/electronics8020122

ADDITIONAL READING

Andreadou, N., Guardiola, M., & Fulli, G. (2016). Telecommunication Technologies for Smart Grid Projects with Focus on Smart Metering Applications. *Energies*, 9(5), 375. doi:10.3390/en9050375

Asad, O., Erol-Kantarci, M., & Mouftah, H. (2013). A Survey of Sensor Web Services for the Smart Grid. *Journal of Sensor and Actuator Networks*, 2(1), 98–108. doi:10.3390/jsan2010098

Khalifa, T., Abdrabou, A., Shaban, K., & Gaouda, A. (2018). Heterogeneous Wireless Networks for Smart Grid Distribution Systems: Advantages and Limitations. *Sensors (Basel)*, 18(5), 1517. doi:10.3390/18051517 PMID:29751633

Lv, S., Kim, H., Zheng, B., & Jin, H. (2018). A Review of Data Mining with Big Data towards Its Applications in the Electronics Industry. *Applied Sciences (Basel, Switzerland)*, 8(4), 582. doi:10.3390/app8040582

Martínez-Álvarez, F., Troncoso, A., Asencio-Cortés, G., & Riquelme, J. (2015). A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting. *Energies*, 8(11), 13162–13193. doi:10.3390/en81112361

Singh, S., & Yassine, A. (2018). Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting. *Energies*, 11(2), 452. doi:10.3390/en11020452

Zhang, J., Hasandka, A., Wei, J., Alam, S., Elgindy, T., Florita, A., & Hodge, B.-M. (2018). Hybrid Communication Architectures for Distributed Smart Grid Applications. *Energies*, 11(4), 871. doi:10.3390/en11040871

KEY TERMS AND DEFINITIONS

Data Mining: A process of extracting information from raw data sets.

Smart Grid: An electrical grid with an integration of power as well as information and communication infrastructure. It is characterized by full duplex flow of energy and information, advanced metering

infrastructure, renewable energy resources, self-healing, active involvement of consumers, and Internet of things.

Wireless Sensor Network: A network comprising of sensor, communication transceivers, memory, processing unit and power supply. Wireless sensor network is useful to convey various parameters such as current, temperature, humidity etc. to the remote monitoring and control unit by using wireless communication technology.

N-Clustering of Text Documents Using Graph Mining Techniques

Bapuji Rao

 <https://orcid.org/0000-0002-2781-9708>

Indira Gandhi Institute of Technology, Sarang, India

INTRODUCTION

Text mining refers to the detection of patterns or similarities in natural language text. For given group of text documents, often the need arises to classify the documents into groups or clusters based on similarity of content. For a small collection, it is possible to manually perform the grouping of documents into specific groups. But to group large volumes of text, the process would be extremely more consumption of time. Therefore, developing a fast and accurate document clustering algorithm is very much helpful for processing of document data sets for clustering.

Clustering is generally studied in data mining problem of text domains. Clustering is an automatic learning technique which aims at grouping a set of objects into Sub-sets or clusters. A clustering can be defined as a grouping of similar objects in the data of similar characteristics. The similarity of characteristics between the objects is measured with the help of a similarity function. The problem of clustering can be very useful in the text domain. Document clustering is extensively used in the areas of text mining and information retrieval. Clustering especially helps of organizing documents in a structural way to improve retrieval and browsing those documents. The study of the clustering problem is related to the applicability to the text domain. Text document clustering is a selection of text documents with the particular word(s)/text(s) present. So each group of text documents called a cluster of text documents of a particular word's presence. Clustering is unsupervised learning it means there is no need of human interference for clustering of documents. In text document clustering, a group of words (texts) is used on a set of text documents for discovering such text documents having with the given set of words (texts). Further such discovered text documents for the given set of words (texts) are grouped into that many clusters of text documents.

Graph mining, is also known as graph-based data mining, is the extraction of meaningful and useful knowledge from a graph representation of data. The most natural form of knowledge that can be extracted from graphs is also a graph. Therefore, the knowledge, sometimes referred to as patterns (or sequence), mined from the data are typically expressed as graphs, which may be sub-graphs of the graphical data.

A Graph $G = (V, E)$ with nodes or vertices in a set of vertices V and a set of edges E . Two nodes $u, v \in V$ are adjacent, if an edge $(u, v) \in E$ exists. The graph representation, that is, a collection of nodes and links between nodes, does support all aspects of the relational data mining process. As one of the most general forms of data representation, the graph easily represents entities, their attributes, and their relationships to other entities. In a bi-partite graph, there are two types of nodes are used. In document clustering, the two types of nodes are document and word. Between the documents and words, the formation of graph is known as document-word bi-partite graph. The relationship between the document-word is the frequency of words present in the related document.

DOI: 10.4018/978-1-7998-3479-3.ch057

BACKGROUND

Document Clustering is also known as Text Clustering is a specific application of text mining and a sub-problem of cluster analyses. If the documents of certain categories are identified and grouped as one cluster, such is called Text Classification Problem. The approach discussed in this chapter is applicable to text documents having characteristics of common word(s) appearances in those text documents for clustering purpose. Finally, each cluster of text documents is represented as a bi-partite graph. In a bi-partite graph, the two types of nodes are document node and word node.

LITERATURE SURVEY

The Scatter-Gather method proposed by authors (Cutting, Karger, Pedersen, & Tukey, 1992) defines the hierarchical organization of documents into coherent categories for systematic browsing of the document collection. It provides a systematic browsing technique with the use of clustered organization of the document collection.

The authors (Aggarwal & Zhai, 2012) define both feature selection and feature transformation methods such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF) are used to improve the quality of the document representation and make it more efficient to text clustering. Feature selection is more common and easy to apply in text clustering in which supervision is available for the feature selection process proposed by authors (Yang & Pedersen, 1997). Since the results of text clustering are highly dependent on document similarity. Such cases the concept of term contributed by authors (Liu, Liu, Chen, & Ma, 2003) is applied. So the contribution of a term can be viewed as its contribution to document similarity.

The technique of concept decomposition uses any standard clustering technique has been studied by authors (Aggarwal & Yu, 2001); (Dhillon, & Modha, 2001) on the original representation of the documents. The frequent terms in the centroids of these clusters are used as basis vectors which are almost orthogonal to one another. The documents can then be represented in a much more concise way in terms of these basis vectors. So the condensed conceptual representation allows for enhanced clustering as well as classification of text documents. Therefore, a second phase of clustering can be applied on this condensed representation in order to cluster the documents much more effectively proposed by author (Salton, 1983). Such a method is tested by authors (Slonim & Tishby, 2000) by using word-clusters in order to represent documents.

The non-negative matrix factorization (NMF) technique is a latent space method, and particularly suitable for clustering of text documents proposed by (Xu, Liu, & Gong, 2003). The NMF scheme is a feature transformation method which is particularly used for clustering of documents. Let A be the $n \times d$ term document matrix. To create k clusters from the underlying documents. Then, the non-negative matrix factorization method attempts to determine the matrices U and V by minimizing the following objective function: $J = (1/2) * \|A - U * V^T\|$. The matrix factorization technique is used to determine word clusters instead of document clusters. Just as the columns of V provide a basis which can be used to discover document clusters, just like the use of columns of U to discover a basis which correspond to word clusters. Since document clusters and word clusters are closely related, and it is often useful to discover both simultaneously, and considered as co-clustering. The co-clustering is proposed by (Dhillon, 2001); (Dhillon & Modha, 2001); (Dhillon, Mallela, & Modha, 2003).

Distance-based clustering algorithms are designed by using a similarity function to measure the closeness between the text objects. The most well known similarity function which is used commonly in the text domain is the cosine similarity function proposed by (Aggarwal & Zhai, 2012). These similarity functions can be used in conjunction with a wide variety of traditional clustering algorithms has been studied by authors (Jain & Dubes, 1998); (Kaufman, & Rousseeuw, 2009).

The k-means clustering algorithm also uses a set of k representatives around which the clusters are built. The simplest form of the k-means approach is to start with a set of k seeds from the original set of documents, and assign documents to these seeds on the basis of closest similarity. In the next iteration, the centroid of the assigned points to each seed is used to replace the seed in the last iteration proposed by authors (Aggarwal & Zhai, 2012).

A graph-based approach to document classification is described. The document sets are represented as graph sets where weighted graph mining algorithm is applied to extract frequent sub-graphs. Further the sub-graphs are processed to produce feature arrays (one per document) for classification. Finally the weighted sub-graph mining is used to ensure classification effectiveness and efficient. So that it is most significant of extracting sub-graphs and proposed by authors (Jiang, Coenen, Sanderson, & Zito, 2010).

The word-clusters by authors (Baker & McCallum, 1998) and co-training methods by authors (Nigam, McCallum, Thrun, & Mitchell, 1998) are used in order to improve the classification accuracy of supervised applications with the use of clustering techniques. To enable an effective clustering process, the word frequencies need to be normalized in terms of their relative frequency of presence in the document and over the entire collection of documents. A common representation used for text processing is the vector-space based TF-IDF representation proposed by author (Salton, 1983). Text clustering algorithms are divided into a variety of different types such as agglomerative clustering algorithms, partitioning algorithms, and standard parametric modelling based methods such as the EM-algorithm. An experimental comparison of different clustering algorithms can be seen in articles (Steinbach, Karypis, & Kumar, 2000); (Zhao & Karypis, 2004). The authors (Yang & Pederson, 1997) state that the feature selection is more common and easy to apply in the problem of text categorization in which supervision is available for the feature selection process. Some examples of such methods are Document Frequency-based Selection, Term Strength, Entropy-based Ranking, and Term Contribution. The authors (Hossain & Angryk, 2007) states that the construction of document-graphs from text documents and applies an Apriori paradigm for discovering frequent sub-graphs from them. It represents the English terms in a hierarchical manner and finally construct the document-graphs using WordNet. Each document can be represented as graph of related terms. Then it searches for frequent sub-graphs using graph mining algorithms. It groups documents in the same cluster even if they do not contain common keywords, but possess the same meaning.

A *text* document is a combination of *document* and *term* and these document and term are proposed by authors (Hand, Mannila, & Smyth, 2001); (Salton & Buckley, 1988).

The authors (Ehsan Hosseini-Asl & Jacek M. Zurada, 2014) have made a comparative study of several non-negative matrix factorizations for document clustering for their performance such as accuracy, entropy, purity and computational complexity and residual mean square root error. The author (Vedanayaki, 2014) has studied various graph mining techniques, tools, and issues.

Clustering of m-text documents based on n-words with (1, 2, ..., n) word(s)-wise using graph mining techniques proposed by authors (Rao & Mishra, 2016). It finally creates n-numbers text documents clusters and is shown every cluster as documents-words bi-partite graph.

The authors (Rao & Mishra, 2017) proposed an algorithm for Clustering of m-text documents based on n-words with individual word-wise using graph mining techniques. It finally creates n-numbers text documents clusters word-wise.

Soft document clustering using a novel graph covering approach is proposed by Dörpinghaus, Schaaf, & Jacobs, (2018). The authors adopted graph-theoretical approach for document clustering and its application on a real-world data set.

The authors (Cadot, Lelu, & Zitt, 2018) surveyed seventeen benchmark clustering methods on text dataset. The authors (Maheswari & Sathiaseelan, 2019) proposed enhanced graph fusion method for document clustering.

Let us consider N documents $D = (d_1, d_2, d_3, \dots, d_N)$ and M terms $T = (t_1, t_2, \dots, t_m)$. Each document d_i can be represented as a vector $V_i = (V_{i,1}, V_{i,2}, \dots, V_{i,M})$ in the M -dimensional space R^M . So $V_{i,j}$ represents a measure of association of the term t_j with the document d_i . The value of $V_{i,j}$ is 0 if the document d_i does not contain the term t_j and otherwise it is a nonzero value which is the indication of frequency. In Boolean representation, $V_{i,j} = 1$ if the term t_j appears in the document d_i ; otherwise $V_{i,j} = 0$. Based on this approach, a text is simply modelled as a *document-term frequency matrix* which is depicted in “Figure 1”. To represent the *document-term frequency matrix* for a set of five documents $\{D_1, D_2, D_3, D_4, D_5\}$ and four terms $\{T_1, T_2, T_3, T_4\}$ is of order 5×4 . Assume the terms are T_1 =Economy, T_2 =Cinema, T_3 =Election, T_4 =Admission. Document D_2 is a vector $(4, 7, 9, 13)$ represents as the second row in the matrix, in which the term Economy appears 4 times, Cinema appears 7 times, Election appears 9 times, and Admission appears 13 times, respectively.

Figure 1.

		Terms			
		T ₁	T ₂	T ₃	T ₄
Documents	D ₁	10	6	8	12
	D ₂	4	7	9	13
	D ₃	20	30	15	40
	D ₄	7	10	3	6
	D ₅	5	8	9	2

This chapter is about the proposal of an algorithm, **n-WDCluster** to cluster of m-text documents with the input of n-words, which produces two sets of clusters namely individual word-wise cluster and (1, 2,, n) word(s)-wise cluster.

PROPOSED METHOD

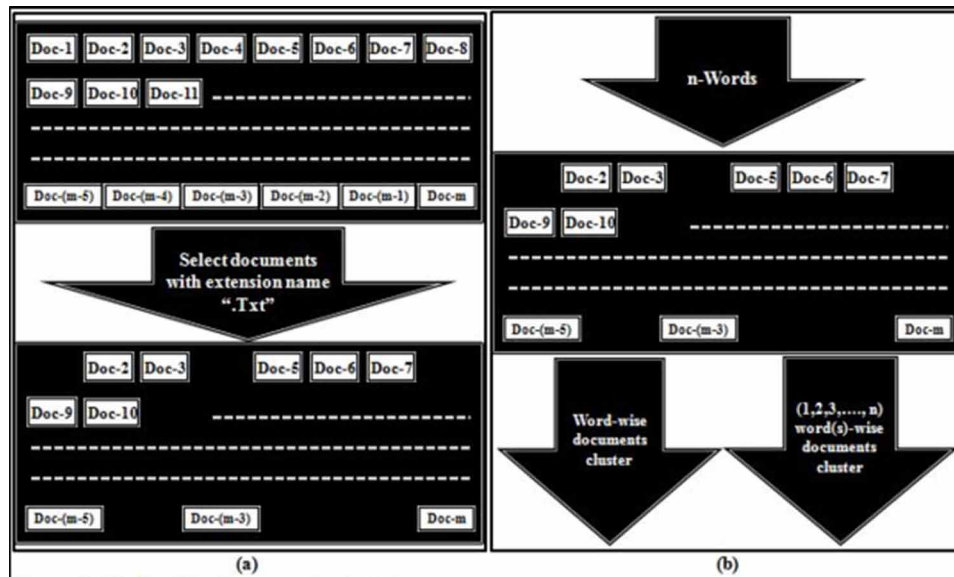
The author propose a method for clustering of only text documents (extension name “.TXT”) word-wise as well as (1, 2,, n) word-wise. A set of m-document names having different extension names are given to the model where it only selects the documents with extension name “.TXT”. The selection of text documents is depicted in “Figure. 2(a)”. On the selected text documents, n-words are applied for

creation of two types of clustering of text documents. The 1st clusters of text documents are formed based on individual word-wise i.e. n -numbers of text documents clusters are formed. Similarly the 2nd clustering of text documents are based on $(1, 2, 3, \dots, n)$ word-wise i.e. again n -numbers of text documents clusters are formed having $(1, 2, 3, \dots, n)$ number of word(s) and depicted in “Figure. 2(b)”.

In case of individual word-wise clusters, the method creates a documents-words frequency matrix of order $m \times n$. By using the frequency matrix, the text documents clustering of individual words are created. Hence there are n -numbers of clustering of documents are formed for the n -numbers of words.

Similarly in case of $(1, 2, \dots, n)$ word-wise clusters of text documents, the documents-words frequency matrix is translated into un-oriented documents-words incidence matrix where all the non-zeros are replaced with 1s. By taking the help of un-oriented documents-words incidence matrix, the $(1, 2, \dots, n)$ word-wise clusters of text documents are formed. Here it too forms n -clusters since there are n -numbers of words. The first cluster of text documents is the presence any one word from n -words. Similarly the n^{th} cluster of text documents is the presence of all n -words in the clustered text documents. Finally these n -clusters are shown as bi-partite graphs since there are two types’ nodes used namely

Figure 2.



words and documents.

PROPOSED ALGORITHM

Algorithm n -WDCluster() [Algorithm Convention (Horowitz, Sahni, & Mehta, 2013).]

// Global Declarations

Document.Txt: Text file to contain 'm' documents name.

Word.Txt: Text file to contain 'n' words.

Document[m]: To assign 'm' text documents name with extension ".Txt".

Word[n]: To assign 'n' words.

Frequency[m][n]: Documents-Words matrix to contain frequencies.

```
{
// to open the file "Document.Txt"
open("Document.Txt");
// to read total number of documents
read(m);
// to pick text documents with extension name ".Txt" and store in the array
Document[ ]
i:=0;
while(not EOF())
do
{
// to read text document name from file "Document.Txt"
read(DocumentName);
Extension:=Get_Extension_Name(DocumentName);
if(Extension=".Txt")
{
i:=i+1;
Document[i]:=DocumentName;
}
}
close("Document.Txt");
// to assign the actual number of ".Txt" documents i.e., 'i' in 'm'
m:=i;
// to open the file "Word.txt"
open("Word.txt");
// to read total number of words
read(n);
i:=1;
while(not EOF())
do
{
// to read words from file "Word.txt"
read(Word[i]);
i:=i+1;
}
close("Word.Txt");
DocWord_MatrixCreation();
// to write the Document-Word Frequency matrix in the file "WIMatrix.Txt"
UIMatrix_Write("WIMatrix.Txt");
// text documents clustering word-wise
```

```
Cluster_Formation_Word();  
// to replace all non-zeros of Frequency[m][n] with 1  
for i:=1 to m do  
{  
  for j:=1 to n do  
    if(Frequency[i][j]≠0) then Frequency[i][j]:=1;  
  }  
// to write the Un-Oriented Documents - Words Incidence Matrix in the file  
"UIMatrix.Txt"  
  UIMatrix_Write("UIMatrix.Txt");  
// text documents clustering (1, 2, ....., n)word-wise  
  Cluster_Formation();  
}
```

PROCEDURE FOR CREATION OF DOCUMENT-WORD MATRIX

Procedure DocWord_MatrixCreation()

flag[m]: To assign the files status i.e. availability or not in the disk

```
{  
  // to check the availability every file names'  
  for i:=1 to m do  
  {  
    flag[i]:=0;  
    // to open the ith text document for reading  
    open(Document[i]);  
    if(Not_Found(Document[i])) then flag[i]:=1;  
    close(Document[i]);  
  }  
  status:=0;  
  for i:=1 to m do  
  {  
    if(flag[i]=1)  
    {  
      write(Document[i],"Not Found");  
      status:=1;  
    }  
    flag[i]:=0; //reassignment  
  }  
  if(status=1) then write("No Text Document Clustering");  
  else // status is OK
```

```

{
  for i:=1 to m do
  {
    // to open the ith text document
    open(Document[i]);
    st:=" ";
    // read a character from ith text document and assign to 'ch'
    read(ch);
    while(not EOF())
    do
    {
      if(ch=space or ch='\n')
      {
        // to check whether the 'st' word is present in Word[] array or
not
        for j:=1 to n do
        {
          if (Word[j]=st) then DocWord[i][j]:=1;
          else DocWord[i][j]:=0;
        }
        st:=" ";
      }
      else st:= st + ch;
      // read character-wise from ith text document and assign to 'ch'
      read(ch);
    } //while close
    close(Document[i]);
  } // for close
} //else close
}

```

PROCEDURE TO WRITE WORD-WISE CLUSTERING OF TEXT DOCUMENTS IN FILES

Procedure Find_Documents_Word(Frequency, Column)

Frequency [m][n]: Documents-Words matrix to contain frequencies.

Column: Column index of word in the matrix Frequency[][].

DIndexWFre[][2]: To assign the detected frequencies document index and its frequency.

```

{
  j:=1;

```



```
for i:=1 to m do
{
    if(Frequency[i][Column]!=0)
    {
        DIndexWFre[j][1]:= i; // document index
        DIndexWFre[j][2]:= Frequency[i][Column]; // word frequency
        j:=j+1;
    }
}
FileName:=Word[Column]+".Txt";
// to open the file 'FileName' for writing purpose
open(FileName);
// to write in the file 'FileName'
Write("Cluster of Documents of Word"+Word[Column]);
for i:=1 to j do
    Write(Document[DIndexWFre[i][1]] + " " +DIndexWFre[i][2]);
// to open the file 'FileName'
close(FileName);
}
```

PROCEDURE FOR CREATION OF WORD-WISE TEXT DOCUMENTS CLUSTERING

Procedure Cluster_Formation_Word ()

```
{
    for i:=1 to n do // n no. of words
        Find_Documents_Word(Frequency, i);
}
```

PROCEDURE TO WRITE (1, 2, 3,, N) WORD(S)-WISE CLUSTERING OF TEXT DOCUMENTS IN FILES

Procedure DetectDocuments(DocWord, words, twords)

DocWord [m][n]: Document-Word Un-Oriented matrix.

words: Actual number of words.

twords: Number of words used for cluster.

Cluster[m]: To assign the text documents index values.

```

{
  for i:=1 to m do
  {
    count:=0;
    for j:=1 to words do
      if(DocWord[i][j]=1) then count:=count+1;
    if(count=twords)
    {
      k:= k+1;
      //to assign text document index
      cluster[k]:=i;
      flag:=1;
    }
  }
  if(flag=1)
  {
    FileName:="CD"+twords+".Txt";
    // to open the file 'FileName' for writing purpose
    open(FileName);
    // to write cluster of documents with 'tword' word(s) presence in the file
    'FileName'
    for i:=1 to nw do
      write(Word[i]);
    // to write text document names and its presence as value 1 in the file
    'FileName'
    for i:=1 to k do
    {
      write(Document[cluster[i]]);
      for j:=1 to nw do
        write(DocWord[Cluster[i]][j]);
      }
    // to close the file 'FileName'
    close(FileName);
  }
  else
    write("No Documents Clustering");
}

```

PROCEDURE FOR CREATION OF (1, 2,....., N) WORD(S)-WISE CLUSTERING OF TEXT DOCUMENTS

Procedure Cluster_Formation ()

```
{  
  for i:=0 to (n-1) do    // n no. of clusters  
  {  
    DetectDocuments(Frequency, n, n-i);  
  }  
}
```

PROCEDURE TO WRITE BOTH WEIGHTED AND UN-ORIENTED INCIDENCE MATRICES IN FILES

Procedure UIMatrix_Write(FileName)

```
{  
  // to open the file 'FileName' for writing purpose  
  open(FileName);  
  
  for i:=1 to n do  
    write(Word[i]); // to write words in 'FileName'  
  for i:=1 to m do  
  {  
    write(Document[i]); // to write text document names in 'FileName'  
    for j:=1 to n do  
      write(Frequency[i][j]); // to write frequencies in 'FileName'  
    }  
    close(FileName); // to close the file 'FileName'  
}
```

PROCEDURE TO VERIFY THE PRESENCE OF WORD IN THE TEXT DOCUMENT

Procedure Verify(Row, String)

Row: Row index of word.

String: Word to verify in the array Word[.].

```
{  
  for i:=1 to n do  
  {  
    st1:=Word[i]+".";  
    st2:=Word[i]+",";"  
    st3:=Word[i]+";";
```

```
// to verify word 'String' of Document[Row] for all words of array Word[]
if(String=Word[i] or String=st1 or String=st2 or String=st3) then
    Frequency[row][i]:=Frequency[row][i]+1;
}
}
```

PROCEDURE TO CREATE DOCUMENTS-WORDS FREQUENCY MATRIX

Procedure DocWord_MatrixCreation()

flag[m]: To assign the files status i.e. availability(0) or not(1) in the disk.

```
{
    // to check the availability of every file name
    for i:=1 to m do
    {
        flag[i]:=0;
        // to open the ith text document for reading purpose
        open(Document[i]);
        if(Not_Found(Document[i])) then flag[i]:=1;
        close(Document[i]);
    }
    status:=0;
    for i:=1 to m do
    {
        if(flag[i]=1)
        {
            display(Document[i],"Not Found");
            status:=1;
        }
        flag[i]:=0; // reassignment
    }
    if(status=1) then display("No Text Document Clustering");
    else // status is OK
    {
        for i:=1 to m do
        {
            // to open Document[i] for reading purpose
            open(Document[i]);
            st:=" ";
            // to read a character from Document[i] and assign to ch
            read(ch);
        }
    }
}
```

```

while(not EOF())
do
{
    if(ch=space or ch='\n')
    {
        // to verify word 'st' in Word[] array
        Verify(i, st);
        st:=""
    }
    else st:= st + ch;
    // to read a character from Document[i] and assign to ch
    read(ch);
} // while
close(Document[i]);
} // for
} // else
}

```

The proposed algorithm, ***n*-WDCluster** has three phases. The Phase-I reads two datasets namely “*Word.Txt*” and “*Document.Txt*”, which contains *n*-words and *m*-text document names respectively. First the dataset file “*Document.Txt*” is opened for reading. It assigns the number of documents to the variable *m* and all *m*-text document names to the array, Document[*m*]. Secondly it opens the second dataset file “*Word.Txt*” for reading purpose. It assigns the number of words to the variable *n* and all *n*-words to the array, Word[*n*]. So Phase-I is all about reading of data i.e. *n*-words and *m*-text document names respectively.

The Phase-II for calling of three procedures namely DocWord_MatrixCreation(), UIMatrix_Write(), and Cluster_Formation_Word(). The 1st procedure is to detect the availability of text documents in the disk. Based on its availability status, it checks all its words with the *n*-words available in the array Word[*n*] by calling the procedure Verify(). If the word of text document is found in the array Word[*n*] then it updates the corresponding word in the matrix Frequency[*m*][*n*]. By calling the 2nd procedure, the frequency matrix is written in the file called “*WIMatrix.Txt*”. The 3rd procedure is called for writing all the *n*-clusters formed word-wise in *n*-text files by calling the procedure Find_Documents_Word(). So Phase-II is about the writing of frequency matrix and *n*-clusters word-wise in (*n*+1) text files.

The Phase-III first converts the frequency matrix Frequency[*m*][*n*] to Boolean matrix by replacing all non-zeros with 1s, which is considered as the un-oriented documents-words incidence matrix. The procedure UIMatrix_Write() is called for writing the Boolean matrix Frequency[*m*][*n*] in the text file called “*UIMatrix.Txt*”. Then the procedure Cluster_Formation() is called which in turn calling the procedure DetectDocuments() for writing (1, 2, 3,, *n*) word-wise cluster in *n*-text files. So Phase-III is about the writing of un-oriented incidence matrix and (1, 2,, *n*) word(s)-wise clusters in (*n*+1) text files.

EXPERIMENTAL RESULTS

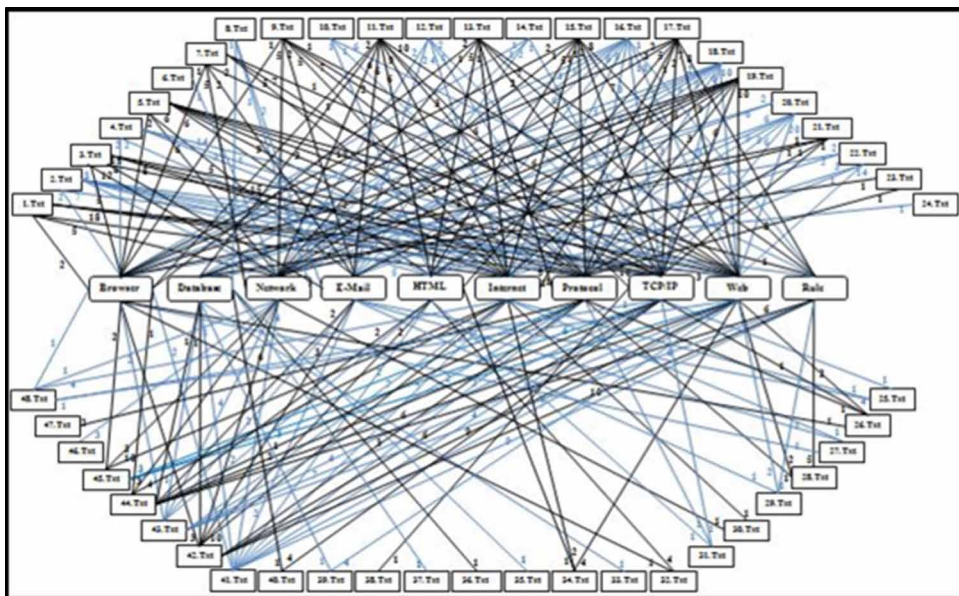
Example

Two dataset files namely “*Document.Txt*” and “*Word.Txt*” are created for the above proposed algorithm ***n*-WDCluster**. In “*Document.Txt*”, the 1st row has the total number of documents and the 2nd rows onwards have the document names. Similarly in “*Word.Txt*”, the 1st row has the total number of words and the 2nd rows onwards have the unique word names. When input these two datasets to the experiment, it opens both the dataset files. The details are read from these two dataset files and assigned to the variable *n*, the array Word[*n*], the variable *m*, and the array Document[*m*] respectively. It reads every document names

Figure 3.

	Browser	Database	Network	E-Mail	HTML	Internet	Protocol	TCP/IP	Web	Rule
1.Txt	2	0	5	0	0	18	5	10	3	0
5.Txt	2	0	6	6	6	9	3	6	10	1
6.Txt	0	0	1	0	0	0	0	0	0	0
2.Txt	2	0	1	0	5	1	1	2	7	8
41.Txt	2	1	4	2	2	7	2	4	7	9
18.Txt	2	1	6	6	6	9	3	6	10	0
45.Txt	3	0	3	0	0	10	3	6	3	0
16.Txt	2	1	3	6	6	9	3	6	10	1
46.Txt	0	0	3	0	0	0	0	0	0	0
9.Txt	1	0	5	2	0	5	7	2	1	0
42.Txt	3	1	6	3	2	10	3	6	9	10
10.Txt	0	0	0	0	1	0	1	0	0	6
14.Txt	1	0	1	2	0	0	0	1	0	0
12.Txt	0	0	2	2	0	4	1	2	0	0
20.Txt	0	2	6	6	8	9	3	6	20	0

Figure 4.



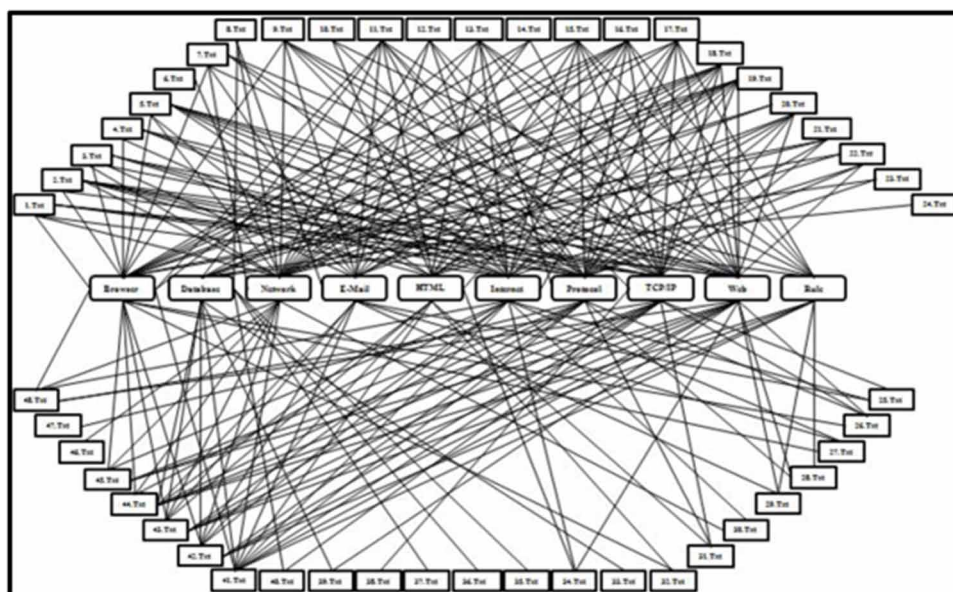
from the array Documents[m] and only selects the document file with extension name “.Txt”. Since the algorithm has to clustering of text documents with extension name “.Txt”. The “Document.Txt” has 74 numbers of documents with different extension names. The algorithm only selects 48 numbers of text documents with extension name “.Txt” and stored in the array Documents[m] having document names 1.Txt, 2.Txt, 3.Txt,....., and 48.Txt respectively. The procedures DocWord_MatrixCreation() and Verify() are responsible of creation of matrix Frequency[m][n]. The procedure UIMatrix_Write() writes the matrix, Frequency[m][n] in a text file called “WIMatrix.Txt” which is considered as the Documents- Words Frequency matrix and depicted in “Figure 3”.

The procdure Cluster_Formation_Word() creates the Documents-Words frequency matrix, Frequency[m][n]. Here the order of Documents-Words frequency matrix is 48x10 since there are 48

Figure 5.

	Browser	Database	Network	E-Mail	HTML	Internet	Protocol	TCP/IP	Web	Rule
1.Txt	1	0	1	0	0	1	1	1	1	0
5.Txt	1	0	1	1	1	1	1	1	1	1
6.Txt	0	0	1	0	0	0	0	0	0	0
2.Txt	1	0	1	0	1	1	1	1	1	1
41.Txt	1	1	1	1	1	1	1	1	1	1
18.Txt	1	1	1	1	1	1	1	1	1	0
45.Txt	1	0	1	0	0	1	1	1	1	0
16.Txt	1	1	1	1	1	1	1	1	1	1
46.Txt	0	0	1	0	0	0	0	0	0	0
9.Txt	1	0	1	1	0	1	1	1	1	0
42.Txt	1	1	1	1	1	1	1	1	1	1
10.Txt	0	0	0	0	1	0	1	0	0	1
14.Txt	1	0	1	1	0	0	0	1	0	0
12.Txt	0	0	1	1	0	1	1	1	0	0
20.Txt	0	1	1	1	1	1	1	1	1	0

Figure 6.



numbers of text documents and 10 numbers of words respectively. Using Documents-Words frequency matrix, $\text{Frequency}[m][n]$, it started creation of word-wise n -clusters of text documents. In this example the 10 words are {Browser, Database, Network, E-Mail, HTML, Internet, Protocol, TCP/IP, Web, Rule}. Hence 10 numbers of clusters are formed and written in text files viz. “Browser.Txt”, “Database.Txt”, “Network.Txt”, “E-Mail.Txt”, “HTML.Txt”, “Internet.Txt”, “Protocol.Txt”, “TCP/IP.Txt”, “Web.Txt”, and “Rule.Txt” respectively.

After formation of word-wise clusters, the frequency matrix, $\text{Frequency}[m][n]$ is converted into Boolean matrix, which is considered as the un-oriented documents-words incidence matrix depicted in “Figure 5”. Then the procedure `Cluster_Formation()` is called which in turn calling the procedure `DetectDocuments()` for creation of 1-word appearance cluster, 2-words appearance cluster, 3-words appearance cluster, 4-words appearance, 5-words appearance, 6-words appearance, 7-words appearance, 8-words appearance, 9-words appearance, 10-words appearance and written in 10 text files viz. “CD1.Txt”, “CD2.Txt”, “CD3.Txt”, “CD4.Txt”, “CD5.Txt”, “CD6.Txt”, “CD7.Txt”, “CD8.Txt”, “CD9.Txt”, and “CD10.Txt” respectively.

Results of Word-Wise Document Clustering

By using “Figure 3”, the author represent graphically as Documents-Words weighted bi-partite graph and depicted in “Figure 4”. There are two types’ of nodes viz. Documents and Words. The edge between the Documents and the Words indicate the frequency of words in the related documents.

Results of (1, 2,....., n) Word(s)-Wise Document Clustering

By using “Figure 5”, the author represents graphically as Documents-Words bi-partite graph and depicted in “Figure 6”. The two types of nodes are Documents and Words. The edge between the Documents and the Words indicate the presence of words in the related documents.

The algorithm was written in C++ and compiled with *DevC++*. The experiment was run on Intel Core I5-3230M CPU + 2.60 GHz Laptop with 4GB memory running MS-Windows 7.

The authors (Dhillon, & Modha, 2001) studied vector space models of large document collections. These models generally have very high-dimensional and sparse, which require unique computational and statistical challenges.

The document clustering using word clusters is a double-clustering procedure which follows two stage processes. The first procedure finds the word-clusters. In second procedure the document clusters are obtained by using the word-clusters is proposed by (Slonim & Tishby, 2000).

The authors (Xu, Liu, & Gong, 2003) proposed a novel document clustering method based on the non-negative factorization of the term document matrix of the given document corpus. The non-negative matrix factorization (NMF) captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic with the document’s largest projection value.

The authors (Rege, Dong, & Fotouhi, 2006) proposed a novel idea of modelling set of documents having similar characteristics as a bipartite graph between documents and words. So that simultaneous clustering of documents and words are formed. Then the authors use a new spectral co-clustering algorithm to partition the co-clustering into individual clusters of documents and words.

The main problem in contingency table analysis is co-clustering i.e. simultaneous clustering exist at rows and columns. The authors (Dhillon, Mallela, & Modha, 2003) proposed a co-clustering algorithm that tries to preserve mutual information by twining both the row and column clustering at all stages.

In text classification using graph mining based feature extraction is proposed by (Jiang, Coenen, Sanderson, & Zito, 2010). It represents nodes as sentence, and the internal structures of noun (NP), verb (VP) and prepositional phrases (PP) are represented by triangles. The part of speech (POS) is represented as circle, token as rectangles, and semantic as ovals. This technique is about to extract frequent sub-graphs.

CONCLUSION

The proposed algorithm, n -WDCluster clusters only text documents after input of m -documents of various extension name. The algorithm only selects documents with extension name “.Txt” from m -documents. These m -text documents are clustered on the input of n -words. The algorithm successfully forms n -numbers of cluster of two formats. The first format of n -clusters is formed word-wise. Similarly the second format of n -clusters is formed $(1, 2, \dots, n)$ word(s)-wise. The proposed algorithm, n -WDCluster was implemented using C++ programming with the datasets depicted in the “Figure 3”. The author has observed satisfactory results of creation of *ten* numbers of word-wise clusters as well as ten numbers of $(1, 2, \dots, 10)$ word(s)-wise clusters of text documents. Here the clustering of document-word is one type of bi-partite graph since two types of nodes such as document and word are used. The algorithm was written in C++ language. The experiment was run on Intel Core I5-3230M CPU + 2.60 GHz Laptop with 4GB memory running MS-Windows 7 and observed satisfactory results.

FUTURE RESEARCH DIRECTION

This proposed algorithm of this chapter clusters text documents with extension name “.TXT” word-wise. The future work lies with clustering of documents containing text and images with any extension name image-wise.

REFERENCES

- Aggarwal, C., & Yu, P. S. (2001). On effective conceptual indexing and similarity search in text data. In *Proceedings of IEEE International Conference on Data Mining, 2001, ICDM 2001*, (pp. 3-10). 10.1109/ICDM.2001.989494
- Aggarwal, C., & Zhai, C. (2012). *A survey of text clustering algorithms*. Mining Text Data. doi:10.1007/978-1-4614-3223-4_4
- Baker, L. D., & McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 96-103). 10.1145/290941.290970
- Cadot, M., Lelu, A. & Zitt, M. (2018). *Benchmarking seventeen clustering methods on a text dataset* [Research Report].

- Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 318-329). 10.1145/133160.133214
- Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 269-274). 10.1145/502512.502550
- Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, (pp. 89-98). 10.1145/956750.956764
- Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2), 143–175. doi:10.1023/A:1007612920971
- Dörpinghaus, J., Schaaf, S. & Jacobs, M. (2018). Soft document clustering using a novel graph covering approach. *BioData Mining*, 11(11), 1-20.
- Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.
- Horowitz, E., Sahni, S., & Mehta, D. (2013). *Fundamentals of Data Structures in C++* (2nd ed.). Universities Press (India) Private Limited.
- Hossain, M. S., & Angryk, R. A. (2007). Gdclust: A graph-based document clustering technique. In *2007 IEEE International Conference on Data Mining (ICDM'07), IEEE ICDM Workshop on Mining Graphs and Complex Structures*, (pp. 417-422). IEEE.
- Jain, A., & Dubes, R. C. (1998). *Algorithms for Clustering Data*. Prentice Hall.
- Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text classification using graph mining-based feature extraction. *Knowledge-Based Systems*, 23(4), 302–308. doi:10.1016/j.knosys.2009.11.010
- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.
- Liu, T., Liu, S., Chen, Z., & Ma, W. Y. (2003). *An evaluation on feature selection for text clustering* (Vol. 3). Icml.
- Maheswari, M.U., & Sathiaselvan, J.G.R. (2019). Text Graph - An Enhanced Graph Fusion Model for Document Clustering. *International Journal of Innovative Technology and Exploring Engineering*, 8(7S2), 640-644.
- Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (1998). Learning to classify text from labeled and unlabeled documents. *AAAI/IAAI*, 792.
- Rao, B., & Mishra, B. K. (2017). An Approach to Clustering of Text Documents Using Graph Mining Technique. *IJRSDA*, (Vol. 4), Issue 1. Article, 5, 38–55.
- Rao, B., & Mishra, S. N. (2016). An Approach to Text Documents Clustering with $\{n, n-1, \dots, 1\}$ -Word(s) Appearance Using Graph Mining Techniques. *IJSEAT*, 4(12), 756-762.

Rege, M., Dong, M., & Fotouhi, F. (2006). Co-clustering documents and words using bipartite isoperimetric graph partitioning. In *Proceedings of Sixth International Conference on Data Mining, ICDM'06*, (pp. 532-541). 10.1109/ICDM.2006.36

Rutkowski, L., & ... (Eds.). (2014). *ICAISC 2014, Part II, LNAI 8468*. Springer International Publishing Switzerland.

Salton, G. (1983). *An Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 208-215). 10.1145/345508.345578

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining*, 400(1), 525-526.

Vedanayaki, M. (2014). Graph Mining Techniques, Tools and Issues - A Study. *Indian Journal of Science and Technology*, 7(S7), 188–190.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 267-273). 10.1145/860435.860485

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization* (Vol. 97). ICML.

Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3), 311–331. doi:10.1023/B:MACH.0000027785.44527.d6

KEY TERMS AND DEFINITIONS

Bi-Partite Graph: A graph with two types of nodes viz. Word node and Document node. The relationship between the word and document is an edge which is the indication of word in the related document.

Cluster: The set of related documents which consist of a particular word.

Documents-Words Frequency Matrix: A matrix which consists of only the words frequency in documents.

Un-Oriented Documents-Words Incidence Matrix: A matrix consists of only 1 which is the indication of presence of word in documents.

Characterization of Elevated Tumor Markers in Diagnosis of HCC Using Data Mining Methods

Vyshali J. Gogi

 <https://orcid.org/0000-0002-2814-567X>

Rashtreeya Vidyalyaya College of Engineering, India

Vijayalakshmi M. N.

Rashtreeya Vidyalyaya College of Engineering, India

INTRODUCTION

The process of identifying patterns in huge datasets comprising methods such as machine learning, statistics, and database system can be considered data mining. It is a multidisciplinary field in computer science, and it excerpts knowledge from the massive data set and converts into comprehensible format. The Medical environment is rich in information but weak in knowledge. Medical systems contain wealth of data which require a dominant analysis tool for determining concealed association and drift in data. The health care condition that comprehends to liver disorder is termed as Liver disease. Liver disorder leads to abrupt health status that precisely governs the working of liver and intern affecting other organs in the body. Data mining classification techniques like Decision Tree, Linear Discriminant, and SVM Fine Gaussian and Logistic Regression algorithms are applied.

BACKGROUND

Hepatocellular Carcinoma (HCC) is an abnormal condition of human liver which is diagnosed at a very advanced stage. The disease is liver disorder which can be predicted after series of clinical and laboratory and imaging studies. Hepatocellular carcinoma is the most malignant tumors which is the major cause of death and requires to be treated at early stage. Benign tumors do not lead to cancer but malignant tumors lead to cancer. The lab investigation accords to austerly of the disease in liver together with physiology of the condition. The laboratory studies incorporate persistence of electrolytes, Complete Blood Count, Liver Function Tests, International Normalized Ratio (INR), partial thromboplastic time (PTT) and AFP (Alpha-Feto Protein) tests. Alpha-fetoprotein (AFP) is considered as one of the tumor markers of HCC. AFP is produced by tumors or by regenerating hepatocytes, but not all tumors produce AFP. The elevations in AFP level greater than 400ng /DL in proper clinical context and appropriate radiologic findings marks the diagnosis of HCC. Increased AFP along with proper imaging studies is considered diagnostic of HCC. The medical imaging techniques like USG (Ultrasonography), CT (Compute Tomography) scan, and MRI (Magnetic Resonance Imaging) creates an interior visualization of the body. Surgical resection or ablation cannot be planned with the anatomic detail obtained from ultrasonography. The lesions are not detected significantly. Although CT scan is highly accurate compared to ultrasonography,

DOI: 10.4018/978-1-7998-3479-3.ch058

it may miss some of small lesions. MRI is an excellent method to characterize HCC. HCC demonstrates various features which is clearly differentiated in MRI scan. Patients with lesions larger than 1 to 2 cm are recommended for biopsy since they are predicted to be under risk of malignancy. The decision on resection, transplantation and ablative therapy is taken based on the biopsy result.

According to the literature survey advanced hepatocellular carcinoma was treated with Sorafenib molecular target drug which increased the survival time of the patient with 3 to 4 months (Gomaa W., 2017). Presence of Hepatitis C virus and non-alcoholic fatty liver also contributed to liver disorder. Researchers developed a framework considering the imaging studies and found that imaging studies alone was not sufficient for efficient diagnosis of the disease. Artificial Immune Systems and Artificial Neural Network algorithms were applied on laboratory parameters. Artificial Neural Network gave the better result (Shane Dixon & Xiao-Hua Yu, 2016). Data mining algorithms support vector machine and back propagation algorithm was applied on the liver function test attributes. Support vector machine gave best results. Micro array analysis was used as molecular biology approach where the result was affected by diet, age, and ethnicity (Sumedh Sontakke, 2017). Clinical decision support systems support physicians to carry out meticulous prediction of the disease. The Volume of the dataset, and attributes present in the dataset along with the competence and deficiency of the classifiers contributes to deal with medical issues (Ahmed J. Aljaaf, 2015). A cancer prediction system provides warning to the patients thereby reducing the cost and saving time. Validation of the system is done by comparing the predicted results with the previously available medical record of the patient (K.Arutchelvan, 2015). The study of hepatitis dataset and its contribution to Hepatocellular carcinoma. Machine learning algorithms along with feature selection algorithms were applied on the hepatitis dataset. K-fold cross validation was implemented where diverse number of fold were applied and performance was analyzed (Tri Astuti, 2015). The use of logistic regression and artificial neural network models for biomedical data classification was presented by the researchers. The study gave the conclusion that no single algorithm can outperforms all other algorithms. The result depends on the kind of data and cases (Stephan Dreiseitla, 2003). Eleven data mining classification algorithms were applied to the datasets and the performance of all classifiers are compared against each other in terms of accuracy, precision, and recall. Several investigations have also been carried out to improve performance of the classification models. Finally, the results shown promising methodology in diagnosing liver disease during the earlier stages (Sina Bahramirad, 2013). Abnormal LFT result not only mean that the problem is pertaining to liver always, it may also indicate abnormalities in other organs. To overcome this, researchers came out with a prediction model called Algorithm for Liver Function Investigations (ALFI), which gave a decision support tool incorporating a clinical prediction rule that facilitated for the management of patients in primary care. The C-Statistic technique was used for the study (David J. McLernon, 2012). A framework was proposed based on ultrasound images for classifying liver fibrosis. Deep features were learnt via transfer learning and Visual Geometry Group neural network (VGGNet). Fully connected network (FCNet) classifier was trained in order to predict the normal, early stage and late stage fibrosis liver status (Dan Meng L. Z., 2017). Each stage of liver cancer is diagnosed depending on extent to which liver is damaged. To determine the degree of liver damage at four different stages, a KNN based intelligent computational method was presented. The accuracy of the classification was compared with linear discriminant analysis, diagonal linear discriminant analysis, quadratic discriminant analysis, diagonal quadratic discriminant analysis classifiers and classification and regression trees (Singh & Pandey, 2016).

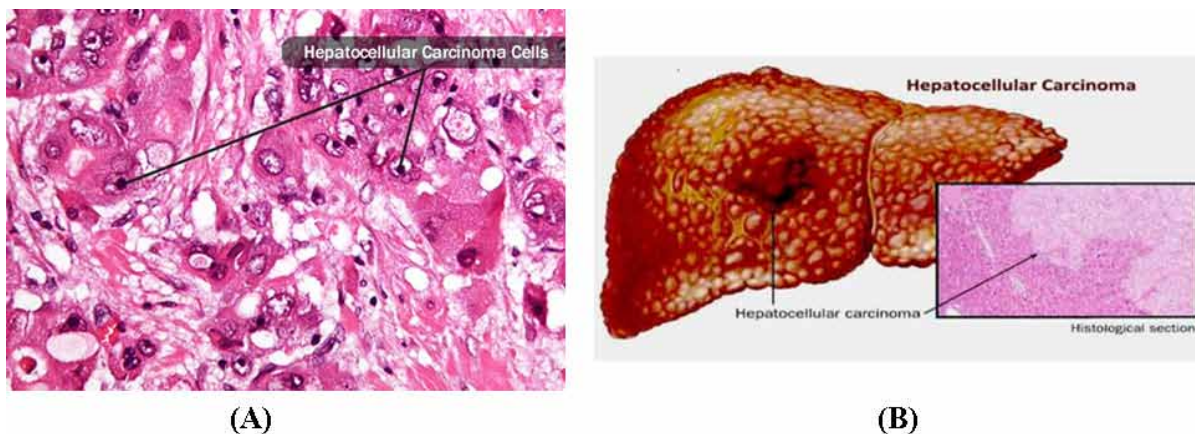
Based on the literature survey, the authors have used both statistical and computational techniques for diagnosis of Hepatocellular Carcinoma. The authors in the computational methods used laboratory parameters and imaging studies, and the authors in statistical techniques considered tumor markers.

Diagnosis of Hepatocellular Carcinoma requires the study of tumor markers along with laboratory parameters. The chapter focusses on study of tumor marker along with liver function test attributes.

FOCUS OF ARTICLE

Healthcare domain generates huge data which is very complex and vast. The data contains many hidden parameters and patterns which is useful in predicting the disease. Data mining helps in recognizing these hidden patterns and arriving at the diagnosis of the disease. In this chapter the authors are concentrating on HCC tumor makers. The aim of the study is to predict the presence of tumor markers and their contribution in HCC Progression using data mining techniques. The presence of tumor is initially detected considering elevated AFP ($>400\text{ng/DL}$) as tumor marker. The tumor marker along with other attributes of Liver Function Test helps in determining the aggressiveness of cancer and response to certain drugs (Hall, 2012). The proposed study concentrates on the contribution of tumor markers in diagnosing the HCC at the earlier stage there by increasing the survival time of the patient. The survival time of the patient is diagnosed based on the elevation in tumor marker. The chapter focusses on application of data mining techniques on the LFT attributes along with tumor marker that contributes for HCC Progression. As the HCC progresses, the tumor spreads in multiple lobules inside the liver in form of hepatocellular cells and further multiplies and grows to form large tumor. The below figure gives shows the existence of hepatocellular cells inside the liver. The histopathological images of the liver make it easy to differentiate the abnormality in the liver and the extent to which liver is damaged due to HCC.

Figure 1. Hepatocellular Carcinoma cells in the liver



The standard threshold values of each attribute that are considered in this chapter is given in below table.

The diagnosis of HCC in the previous studies is based on the blood test results, ultrasound imaging, and genomic expression. The imaging studies like CT scan, Ultrasound scanning were not sufficient in identifying the number of lesions in the liver (Dan Meng L. Z., 2017). The Prognosis of liver disease is the reflection of both tumor characteristics like tumor size, location, and tumor biology along with the degree of underlying resection. Statistical analysis techniques are been used by the researchers and they

Table 1. Threshold values of LFT attributes

Attributes	Minimum – Maximum Value
ALP	15-129 IU/L
SGOT_AST	5-34 units
SGPT_ALT	0-40 Units
TP	6.0-8.3 gm/dl
Serum_ALB	3.2-5.0 gm/dl
Serum_GLB	2.3-3.5 gm/dl
AG_Ratio	1.5 - 2.5 gm/dl
Total_Bilirubin	0.2-1.0mg/dl
GGT	15.0-85.0 IU/L
Direct_Bilirubin	0.0-0.2 mg/dl
Indirect_Bilirubin	up to 0.6mg/dl
SerumCreatinineCT	Male: 0.7-1.2, Female: 0.5-0.9
AFP	<10ng/mL

have not yielded an efficient prediction due to various reasons, but the accuracy of their results was not measured (Heba Ayeldeen O. S., 2015). The elevation in tumor markers contributes to predicting the survival time of the patient. The chapter focuses on contribution of tumor markers in diagnosing HCC at earlier stage. The tumor marker test Alpha -Feto Protein test is considered to analyze the presence of HCC.

SOLUTIONS AND RECOMMENDATIONS

Hepatocellular Carcinoma is diagnosed only in its advance stage. The diagnosis requires both bio-chemistry test and tumor marker test. The bio-chemistry test is the liver function test (LFT) that is done out of blood sample. The deviation in parameters of LFT along with the tumor marker confirms the presence of HCC. The stage of HCC is determined based on size of liver, number of lesions found and their size. In the proposed work, the authors consider the tumor marker Alpha-Feto Protein (AFP) level along with the LFT attributes to predict the presence of HCC. The dataset has attributes such as, Age, Gender, ALP (Alkaline Phosphate), Serum_GLB (Serum Globulin), SGOT_AST (Serum glutamic oxaloacetic transaminase_aspartate aminotransferase), Serum_ALB (Serum Albumin), SGPT_ALT (Serum glutamic pyruvic transaminase_alanine aminotransferase), TP (Total Protein), AG_Ratio (Albumin Globulin Ratio), Total_Bilirubin, GGT (gamma-glutamyl transferase), Direct_Bilirubin, Indirect_Bilirubin, SerumCreatinineCT, and AFP (Alfa-Feto Protein).

The dataset consists of details of 250 patients who are predicted of Hepatocellular Carcinoma (HCC). The data is collected from one of the oncology centers. The dataset is divided into 70% training set and 30% test set.

Data mining classification analysis techniques like Decision Tree, Logistic Regression, Naïve Bayes and Support Vector Machines (SVM) algorithms are applied on the dataset. The deviation from normal values of SGPT_ALT, SGOT_AST along with raised values in AFP along with the CT scanning resulted in prediction of Hepatocellular Carcinoma. SGPT_ALT correlates with explicit hepatocellular disease.

ALP correlates to tumors and biliary diseases. The elevated values of ALT, AST and AFP is considered as markers for diagnosing HCC progression.

The decision tree classifier gave the highest accuracy of 0.91, while Logistic Regression, Naïve Bayes and SVM gave 0.81, 0.90, and 0.89 respectively. Data mining R tool is used for the analysis. The following table gives the performance of algorithms with respect to accuracy and other factors.

Table 2. Performance of Algorithms on the dataset

Algorithm	Accuracy in %	No. of True Positives	No. of True Negatives	No. of False Positives	No. of False Negatives	Misclassification Rate
Decision Tree	91	89	2	0	9	0.09
Naïve Bayes	90	89	1	0	10	0.1
SVM	89	89	0	0	11	0.11
Logistic Regression	81	80	1	9	10	0.19

The sensitivity is when the result is actually yes and when it is predicted yes, specificity is when actually no and predicted no. The dataset contains many false positives which means that the patients actually does not have disease but its predicted yes which may cost money and time for patients. The false negatives predicts no when the patient actually have the disease which may cost patient life. Decision Tree gave the maximum accuracy, zero false negatives and low misclassification rate compared to Naïve Bayes, SVM and Logistic Regression.

Figure 2. Confusion Matrix and Statistics obtained from Decision Tree algorithm

```
> print(confusionMatrix(dec_pred, testd$HCC))
Confusion Matrix and Statistics

      Reference
Prediction 0  1
      0  2  0
      1  9 89

      Accuracy : 0.91
      95% CI : (0.836, 0.958)
      No Information Rate : 0.89
      P-Value [Acc > NIR] : 0.327753

      Kappa : 0.2834
      Mcnemar's Test P-Value : 0.007661

      Sensitivity : 0.1818
      Specificity : 1.0000
      Pos Pred value : 1.0000
      Neg Pred value : 0.9082
      Prevalence : 0.1100
      Detection Rate : 0.0200
      Detection Prevalence : 0.0200
      Balanced Accuracy : 0.5909

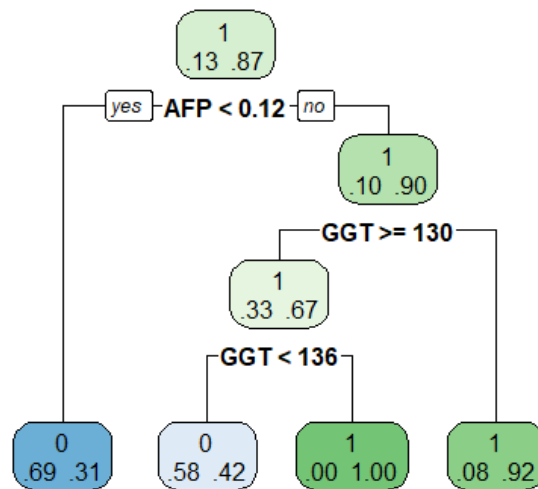
      'Positive' class : 0

> |
```

The figure 2 gives confusion matrix obtained from dataset. Decision tree obtained while plotting the tumor marker AFP is given in figure 3.

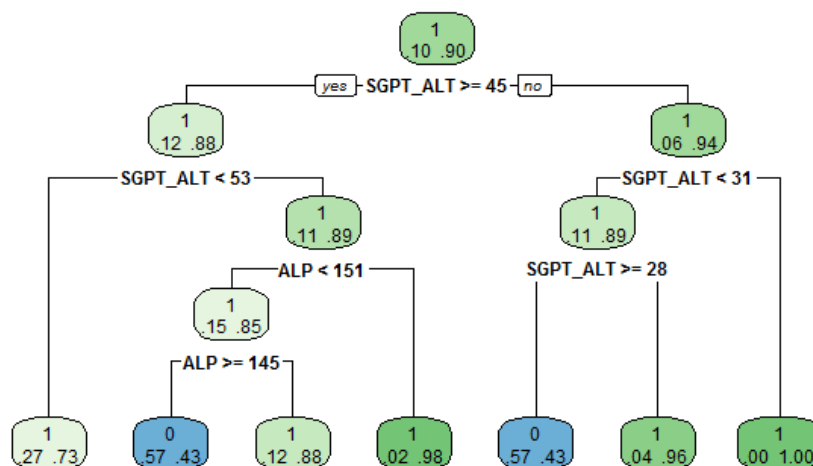
Decision Tree

Figure 3. Decision tree showing AFP



The above figure 3 shows decision tree with tumor marker along with LFT attributes. Decision tree CART is used for the analysis since the dataset contains both numerical and categorical values. The Decision tree formed explains that the tumor marker AFP is more significant than other attributes, along GGT level.

Figure 4. Decision tree showing all attributes except AFP



Decision Tree

The above figure shows the decision tree when tumor marker AFP is excluded. SGPT_ALT is the next significant attribute after AFP to diagnose HCC.

The elevation of AFP along with the LFT attributes ALT and ALP contributed in diagnosis of HCC. The attributes in the dataset ALP, SGOT_AST, SGPT_ALT, are the enzymes present in liver and are considered to be significant parameters in predicting Liver disorder. The data mining algorithms, Decision Tree, Naïve Bayes, SVM and Logistic regression were applied on the dataset. Decision tree gave the highest accuracy 91% with sensitivity 0.9 and specificity equal to 1.

FUTURE RESEARCH DIRECTIONS

Further work can be carried out by considering the parameters from the imaging studies and other confirmative tests. The confirmative test such as FNAC (Fine Needle Aspiration Cytology) along with the AFP test can confirm the presence of HCC and the stage of the disease. The diagnosis of HCC also includes other tumor markers such as DCP (Des-gamma-carboxy prothrombin), GP73 (Golgi protein 73) and AFP-L3. These markers can be further explored to arrive at accurate prediction. More research can be carried out by applying various machine learning algorithms.

CONCLUSION

Hepatocellular carcinoma being one of the deadly diseases and major cause of death world-wide is diagnosed in its advance stages most of the time. The research work was carried for development of organized structure to find out the presence/ absence of HCC disease among the patients using the reports and symptoms collected from them. The generated data was structured and stored in the necessary format, clinical and operational data was configured. The work helped in finding strengths and weakness of various previous work and building focus on unexplored parameters of HCC. The prediction of disease involves various tests and analysis considering various parameters. The authors have considered Liver Function Test along with the tumor marker AFP. The researchers found that elevation of AFP along with the LFT attributes ALT and ALP contributed in diagnosis of HCC. Decision Tree, Naïve Bayes, SVM and Logistic regression are applied on the dataset. Decision tree gave the highest accuracy 91% with sensitivity 0.9 and specificity equal to 1.

REFERENCES

- Aljaaf, A. J., Al-Jumeily, D., Hussain, A. J., Fergus, P., Al-Jumaily, M., & Radi, N. (2015, September). Applied machine learning classifiers for medical applications: clarifying the behavioral patterns using a variety of datasets. In *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 228-232). IEEE. 10.1109/IWSSIP.2015.7314218
- Arutchelvan, K., & Periyasamy, R. (2015). Cancer Prediction System Using Data Mining Techniques. *Int. Res. J. Eng. Technol*, 2(8), 1179–1183.

- Astuti, T., & Adiugroho, H. (2015). The Impact of Different Fold for Cross Validation of Missing Values Imputation Method on Hepatitis dataset. *IEEE Conference Publication*. 10.1109/QiR.2015.7374894
- Ayeldeen, H., Shaker, O., & Ayeldeen, G. (2015). Prediction of liver fibrosis stages by machine learning model: A decision tree approach. *IEEE Conference Publication*. 10.1109/ICoCS.2015.7483212
- Bahramirad, S., & Mustapha, A. (2013). Classification of liver disease diagnosis: A comparative study. *IEEE Conference Publication*. 10.1109/ICoIA.2013.6650227
- Beste, L.A. (2015). *Trends in burden of cirrhosis and hepatocellular carcinoma by underlying liver disease in US veterans*. Academic Press.
- Dixon, S., & Yu, X.-H. (2016). Liver disorder detection based on artificial immune systems. *IEEE Conference Publication*.
- Dreiseitla, S., & Ohno-Machado, L. (2003). *Logistic regression and artificial neural network classification models: a methodology review*. Academic Press.
- El-Halees, A. M., & Shurab, A. H. (2017). Blood tumor prediction using data mining techniques. *Blood Tumor Prediction Using Data Mining Techniques*, 6.
- Gomaa, A., & Waked, I. (2017). Management of advanced hepatocellular carcinoma: review of current and potential therapies. *Practice*, 17, 18.
- Hall, P., & Cash J. (2012). *What is the Real Function of the Liver Function Tests?* Academic Press.
- Karthick, R., & Malathi, A. (2015). Preprocessing of Various Data Sets Using Different Classification Algorithms for Evolutionary Programming. *IEEE Conference Publication*.
- Kim, S., & Jung, S. (2014). Effective Liver Cancer Diagnosis Method based on Machine Learning Algorithm. *IEEE Conference Proceedings*. 10.1109/BMEI.2014.7002866
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17. doi:10.1016/j.csbj.2014.11.005 PMID:25750696
- Kurt, I., & Ture, M. (2006). *Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease*. Elsevier Ltd.
- McLernon, D. J., & Dillon, J. F. (2012). *The Utility of Liver Function Tests for Mortality Prediction within One Year in Primary Care Using the Algorithm for Liver Function Investigations*. ALFI. doi:10.1371/journal.pone.0050965
- Meng, D., & Zhang, L. (2017). Liver Fibrosis Classification Based on Transfer Learning and FCNet for Ultrasound Images. *IEEE Journals & Magazine*.
- Qin, Q. F. (2017). *Combination of serum tumor markers dickkopf-1, DCP and AFP for the diagnosis of primary hepatocellular carcinoma*. Elsevier. doi:10.1016/j.apjtm.2017.03.016
- Saha, S., & Sankhadip. (2015). Classifier fusion for liver function test based Indian jaundice classification. *IEEE Conference Publication*. 10.1109/MAMI.2015.7456588

Singh, A., & Pandey, B. (2016). An euclidean distance based KNN computational method for assessing degree of liver damage. *IEEE Conference Publication*. 10.1109/INVENTIVE.2016.7823222

Singh, A., & Pandey, B. (2016). *Classification of primary biliary cirrhosis using hybridization of dimensionality reduction and machine learning methods*. Academic Press.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Jingshen Yixue*, 27(2), 130. PMID:26120265

Sontakke, S., & Lohokare, J. (2017). Diagnosis of Liver Diseases using. *Machine Learning*.

Sweidan, A. H., & El-Bendary, N. (2014). Machine Learning based Approach for Water pollution detection via fish liver microscopic images analysis. *IEEE Conference Publication*. 10.1109/ICCES.2014.7030968

KEY TERMS AND DEFINITIONS

AFP: Alpha feto protein (AFP) is one of the standard tumor markers in detection of Hepatocellular Carcinoma. The test is done by drawing blood from the patient and measuring the AFP levels.

Benign Tumors: A tumor is an abnormal mass of tissue. Benign tumors are not cancerous and cannot spread. A benign tumor remains constant and do not grow, and they do not return after removing.

Confusion Matrix: The performance of the classification model is given by confusion matrix in the form of table. The dataset is a test set whose true values are known.

CT Scan: Computed tomography (CT) scan is a type of imaging study of human body. Computers and rotating X-Ray machines provides comprehensive information regarding the bold vessels, bones in various parts of body and blood vessels.

FNAC: A symptomatic method where a small chunk of tissue is sucked by inserting a needle into body and examined with the help of microscope is called FNAC (Fine Needle Aspiration Cytology). FNAC is a diagnostic procedure where a needle is inserted into the body and a small amount of tissue is sucked for examination under a microscope. If any suspicious lump is found, or if any abnormality is detected in imaging tests, the FNAC is performed.

LFT: Liver function test (LFT) is a bio-chemistry test which is done to determine the health of the liver by measuring various enzymes present in the blood.

Malignant Tumors: Malignant tumors are unwanted mass of tissue, which is cancerous. It spreads, multiplies and gets worse damaging the organs.

MRI: The MRI (magnetic resonance imaging) creates a structure and cross- sectional image of the internal organs using a large magnet, radio waves and computer.

Screening: Early diagnosis of any disease is essential to avoid worst situations. Screening is a preventive measure to identify probable health problem without having any symptoms or sign of the disease.

Sensitivity: If the test diagnoses the patient with disease as having the disease, then it is referred to as sensitivity.

Specificity: If the test diagnoses the patient without disease as not having disease, then it is referred to as specificity.

Ultrasonography: The motion of the organs and structures in the body are captured briskly using ultrasound images. The growth of alien objects in the organs like breasts, thyroid gland, limbs and lymph nodes can be recognized using ultrasound images.

Designing a Concept–Mining Model for the Extraction of Medical Information in Spanish

Olga Acosta

Singularyta SpA, Chile

César Aguilar

 <https://orcid.org/0000-0003-1940-9933>

Pontificia Universidad Católica de Chile, Chile

INTRODUCTION

In recent years, the automatic processing of biomedical information has been benefited from advances made by data and text mining. An example of this advance is the book edited by Ananiadou & McNaught (2006), who give a special relevance to create and use tools capable to extract information from large collections of documents, particularly **PubMed** (www.ncbi.nlm.nih.gov/pubmed/). In fact, given this dimension, Pubmed is the most important repository for obtaining biomedical data, what has motivated the generation of different projects related to computational linguistics such as the **Corpus Genia** (www.geniaproject.org), the **MEDIE** search engine (www.nactem.ac.uk/medie/), or the **Open Biological and Biomedical Ontology Project** (<http://obofoundry.org/>), focused on the development of ontologies that provide an organized knowledge system in biomedicine.

In line with these projects, it is exposed here a method for performing a concept mining on biomedical documents in Spanish. This model is based on the automatic extraction of definitional contexts (or DCs), according to the framework developed by Sierra *et al.* (2008), Sierra (2009), Aguilar & Acosta (2016), as well as Aguilar *et al.* (2016). Our model has been sketched having in mind the following objectives:

- The linguistic analysis of definitional contexts identified in biomedical texts in Spanish.
- The creation of a linguistic corpus in Spanish that is representative for the biomedical area.
- The use of stochastic methods in order to provide empirical evidence to validate in linguistic analysis previously performed.

These objectives allow to establish a set of concrete tasks that can be implemented as modules for a mining system, concretely:

- a) Extraction of terminological information: this one focuses on establishing a chain of text processing, which considers: (i) the selection, tokenization and syntactic annotation of a Spanish corpus in medicine; (ii) the identification of uni/multiword terms using a hybrid method (Acosta, Aguilar & Infante, 2015).
- b) Identification of lexical relations: taking advantage of the term extraction and linking them to their definitions, it is possible to implement an ontology in Spanish, considering the recognition

DOI: 10.4018/978-1-7998-3479-3.ch059

of lexical-semantic relations, specifically hyponymy-hyperonymy and meronymy, based on the proposal of Arp, Smith & Spear (2015).

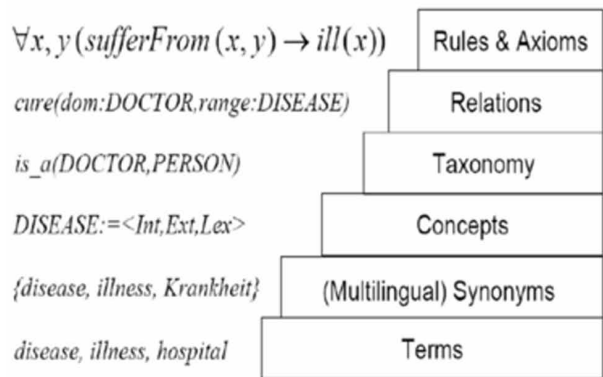
The organization of this article is as follows: section 1 shows a general background behind the notion of concept mining; the section 2 describes how is performed the identification of DCs, in order to detect and extract terms that function as hyponyms, hypernyms and meronyms. The section 3 exposes the results of these extractions, delineating a possible ontology that organizes such lexical-semantic information. To conclude, the section 4 offers a summary along with a brief discussion respect to future applications of this model of mining concepts.

BACKGROUND

Due to the current increasing amount of biomedical literature available on the Web, there is an interest for developing lexical-semantics resources oriented to improve searching and classification of biomedical concepts expressed in large-collections of texts. According to Sainani (2008) biomedicine has been characterized as an area with constant textual production, not only because of the large number of experts working within it, but also because today’s technological progress allows it to discover new knowledge, which must necessarily be documented and diffused in an expeditious and precise manner. An example is **MedLine** (<https://medlineplus.gov/>), a bibliographic database, which has an approximate total of 27.3 million citations and references to articles from 4,500 medical journals published in the United States and in more than 70 countries (Ananiadou, Kell and Tsujii, 2006). For accessing to such a MedLine, it is necessary to use the query engine PubMed (www.ncbi.nlm.nih.gov/pubmed/). Both resources, MedLine and PubMed, have a relevant impact in tasks related to the mining of data and texts.

Other example of this interest is the application to artificial intelligence system like IBM Watson (www.ibm.com/watson/health) for solving automatic analysis of medical data. Such analysis considers the use of natural language processing (NLP) techniques for inferring relevant terms and named entities that refer diseases, in order to support clinical diagnostics (Hoyt *et al*, 2016; Kaggal *et al*, 2016; Boukenze, Mousannif & Haqiq, 2016).

Figure 1. Development of an ontology according to the extraction of textual patterns (Buitelaar, Cimini and Magnini, 2005)



In this regard, Zweigenbaum *et al* (2007) explain that text mining has been applied in biomedicine to identify terms, named entities and lexical-semantic relationships. Furthermore, according to Bichindaritz & Akkineni (2006), as well as Poesio & Almuhareb (2008), the integration of these objects allows the configuration of semantic attributes, which configure concepts. From this approach, the concepts can be recognized in a document, in such a way that one can think of an extraction method capable of identifying them in specific textual patterns. Linking text mining with the creation of ontologies, Buitelaar, Cimiano & Magnani (2005) propose an extraction method that, in a staggered way, identifies such patterns:

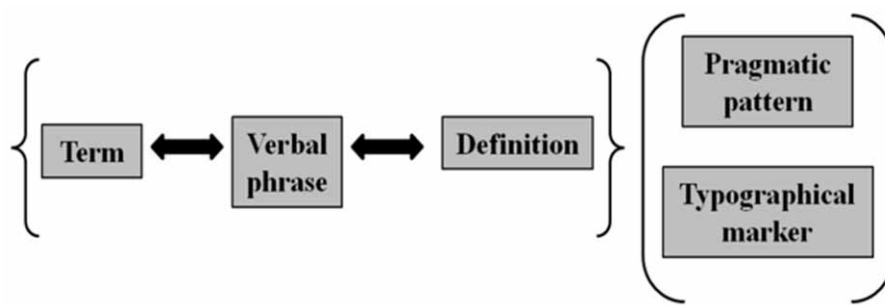
According to Figure 1, in the first step a term extraction in a corpus is conceived. The next step is to search for possible synonyms linked to the terms extracted, either from the query of dictionaries, or using lexical networks such as WordNet. After this task, the next is to build a taxonomy, establishing a link among concepts with terms, according to a lexical relationships between hypernyms, hyponyms, meronyms, holonyms, etc. Then, in a plane of semantic representation, it is possible to delimit non-hierarchical relationships, in order to conceive semantic frames (e. g.: Baker, Fillmore & Lowe, 1998; Baker, Fillmore & Cronin, 2003), capable to shape a richer conceptual scheme. Finally, the last step represents a logical deduction for sketching the axioms that build an ontology, taking into account the conceptual relationships previously inferred (that is, the union among the lexical hierarchy whit the semantic frames).

DEFINITIONAL CONTEXT: A BRIEF DESCRIPTION

Following the experiments conducted by Navigli & Velardi (2004), as well as Velardi, Faralli & Navigli (2013), concept mining takes advantage using linguistic criteria to design algorithms capable to identify lexical and syntactic patterns associated to definitions. These authors have coincidences with the method proposed by Buitelaar, Cimiano and Magnini (2005) when designing **OntoLearn**, a system focused on the construction of ontologies based on the mining of terms and definitions. This system, it should be noted, is directed towards the processing of texts in English.

In parallel, Sierra, Alarcón & Aguilar (2006), Sierra *et al* (2008), as well as Sierra (2009) have been delineating a method of extracting concepts for Spanish through the identification of Definitional Contexts (henceforth, DCs). *Grosso modo*, a DC is text structure that introduces a term linked to a definition. There is a set of connectors that allows to establish this link, such as verbal phrases, pragmatic patterns and typographic marks, among others. See the following figure for more details:

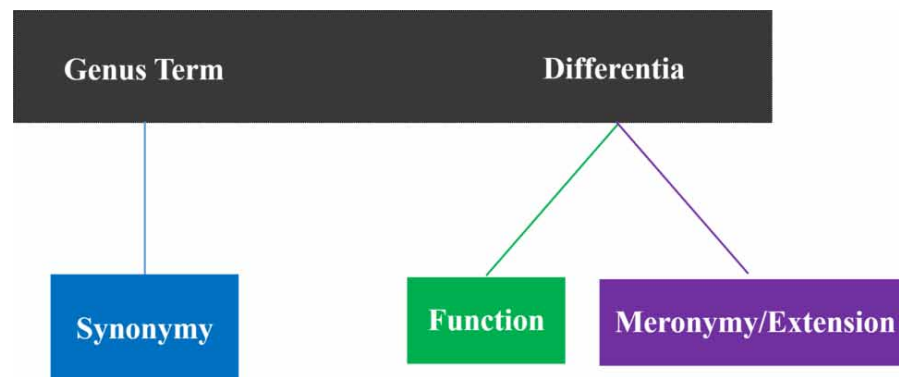
Figure 2. Constitutive units of a DC structure



The Figure 2 shows the configuration of a DC around its basic components (term, verbal phrase and definition), as well as optional units (pragmatic patterns and typographic markers). The combination of these elements configure the linguistic structure of a DC within a text, in such a way that its detection helps to extract both terms and definitions directly.

A relevant subject that emerges in the light of the experiments carried out in the extraction of DCs, is to observe that the definitions assume specific features to express the concept associated with a term, depending on the type of verb that appears in the verbal predication. These features have to do with the formulation of the two basic components of the analytical definition conceived by Aristotle: *Genus Term* and *Differentia*. Taking into account the Aristotle's proposal, Sierra *et al* (2008) propose a typology, which distinguishes three derivable definitions from the analytical model, depending on the fact that the *Genus Term* or the *Differentia* are made explicit or not. It is possible to schematize such typology in the following way:

Figure 3. Typology of definitions according to the Aristotelian model (Sierra *et al*, 2008)



This typology can be better schematized in the following way:

- **Analytical or Aristotelian Definition:** will occur when the *Genus Term* and the *Differentia* appear explicitly within a definition.
- **Synonymical Definition:** will occur when a definition is made explicit the *Genus Term*, establishing a conceptual equivalence with the term that is defined.
- **Functional definition:** it will be given when the *Differentia* is made explicit, offering a definition of a term from its use or application in a given situation.
- **Extensional definition:** will occur when the *Differentia* is made explicit, which presents a definition that lists the components that make up an object represented by the term to be defined. This list of components will follow an order based on relationships that go from one whole to the parties, or from the parties to the whole.

These 4 types, in addition to expressing conceptual information different from one another, also provide the possibility of establishing a conceptual hierarchy that can be reinforced from the corroboration of existing lexical relationships, specifically: hyponymy/hypernym, meronymy and synonymy.

Analyzing now the relationship between verbal phrases and definitions, since Aristotle it has been considered as a logical operator, and it received the name of *predication* (Bäck, 2000). In fact, the idea

that the predication is linked to definitions has recently been reevaluated from a computational point of view (Guarino 1995, Cocchiarella 2003, Smith, 2004).

In concordance with this revaluation, Sierra *et al* (2008) and Aguilar & Sierra (2009) consider the predication as syntactic structure, which can be represented in a formal way using the predication theory (or PT). This theory is a model derived from Government & Binding Grammar formulated by Chomsky (1981). PT states that all predications indicate a semantic relationship between an entity and a particular property or characteristic feature (Rothstein, 1983; Bowers 1993, 2001; den Dikken, 2006). Syntactically, a predication configures a specific kind of phrase structure, just called *predicative phrase* (PrP), which is mapped by a functional head identified as the core of the predication. A graphical tree representation of a PrP is:

Figure 4. Tree representation for PrP, according to Bowers (1993)

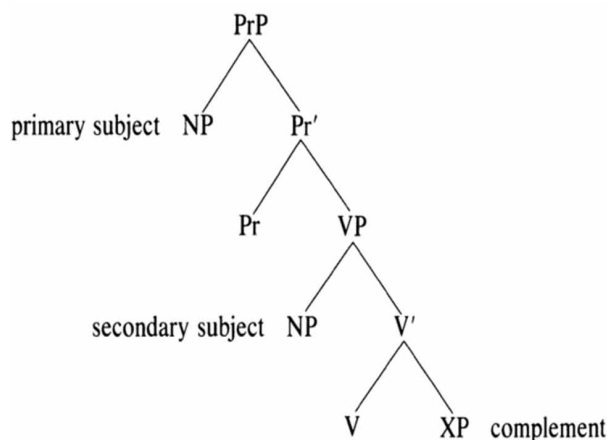


Figure 4 shows the basic configuration of a PrP. According to Bowers (1993, 2001) the head assumes the features [+/-predication] to the functional head. Such head maps two subjects, a primary subject in the position of Specifier of PrP (represented by a Noun Phrase or NP); and a secondary subject, in the position of Specifier of Verbal Phrase or VP (often a NP). Finally, both subjects, the VP and the PrP are linked to one or more complements, which assume many phrasal representations such as NP, IP (Inflection Phrase), CP (Complement Phrase), and so on.

Bowers observes that the VP maps a symmetrical phrase structure to PrP. Therefore, it is possible to consider the subject of a VP as an object associated to the head of a PrP. This association is valid according to the rule of predicate linking. Thus, Bowers conceives two kinds of PrPs:

- a) If the predicate is linked to the subject mapped by the PrP, then it is a primary predication.
- b) In contrast, if the predicate is associated with the subject mapped by the VP, then it is a secondary predication.

We illustrate this distinction in the following examples:

- (i) [A **computer** [_{PrP}] [_{VP}] [_{IP}]].

- (ii) [Turing [defines a **computer** [as a kind of electronic device that processes a set of data_{PrP}] _{VP}] _{IP}]].

In (i), we observe a canonical case of primary predication, where the verb *is* inserts the predicate *a kind of electronic machine...*, linking it to the NP *a computer*. For the term extraction, the noun *computer* is equivalent to a term, while the predicate formulates the definition related to the term.

With regarding to (ii), the verb *defines* maps a PrP composed by three elements: a subject represented by the NP *Turing*, an object expressed by the NP *a computer*, and the predicate *as a kind of electronic device...* This predicate introduces a set of attributes that characterizes *a computer*, following the rule of association *Y is a predicate of X*. Both the predicate and the object configure a predicate linked to the subject through the verb *defines*, generating two symmetrical predicate structures: the first predication covers the subject *Turing*, the main verb, the object and the PrP; in contrast, the second predication covers the object *a computer* with the PrP whose head is the preposition *as*, according with the explanation offered by Bowers (1993, 2001).

From a terminological point of view, it can be considered that the subject of primary predication, as well as the object of secondary predication, represents a **term**. In contrast, the predicate introduced by the PrP represents the definition associated with the term.

FIRST TASK: TERM EXTRACTION

Once the grammatical structure of a DC has been analyzed, the next step has been to perform a term extraction task. For accomplishing this, it was designed a system with a hybrid approach —combining linguistic and stochastic criteria—, which uses a large reference corpus that facilitates the identification of candidates to relevant terms. For determining this assessment, a criterion of relative frequency ratios was used (Manning & Schütze, 2000).

The use of a set of reference texts is an effective method to give a relevant value to the vocabulary recognizable in a domain corpus, considering that the distribution of true and false terms varies between general documents compared to one of specialized texts (Kit & Liu, 2008). Thus, with this measure, a significant portion of the specialized vocabulary will have a high weight, and, consequently, the term candidate segments —specifically, NPs— will be more easily identified in a domain corpus.

The notion of *termhood*, proposed by Kageura & Umino (1996), allows determine the degree of specialization that a term candidate has. Therefore, a contrast is established with those functional words (articles, prepositions, pronouns, etc.), nouns and noun phrases whose meaning does not have any specialization use. Likewise —given its relationship with the notion of *termhood*— the *C-Value/NC-Value* algorithm has been implemented, developed by Frantzi, Ananiadou & Mima (2000). This algorithm works in two phases:

- a) A linguistic, which encompasses three processes: the use of morphosyntactic labeling, the application of nominal sentence formation rules that code potential terms, as well as the use of an unemployment list, which groups together the greatest number of words they do not have a terminological value, in such a way that they are outside the search process.
- b) A statistic, which establishes a measure of *termhood* considering the combination of the following factors: (i) the total frequency of occurrences of a term candidate within a corpus (be it a noun or a nominal phrase); (ii) the total frequency of occurrences of such candidate within a more extensive

string of words; (iii) the total number of extended chains that contain a term candidate; and (iv) the length of words of such chains.

Based on the performance of these phases, the method proposed here allows to set a measure of *termhood* to recognize good candidates. Such a measure is calculated considering formula:

$$\text{Termhood}(a) = f(a) \quad (1)$$

where a represents a string of candidate words at term, and $f(a)$ is its occurrence frequency within a corpus.

In order to better determine this measure, this algorithm also takes into account the neighboring words that appear around a possible term candidate. In short, it is established a contrast among the value of a word can have as a term given its frequency of appearance (*C-Value*), with the degree of specialization that said word has, taking into account its relationship with other neighbors within a text.

This algorithm has been adapted to Spanish by Barrón *et al* (2009), considering some modifications, namely: (i) the delimitation of a set of rules and syntactic patterns that operate as filters for the selection of nouns and nominal phrases that be candidates for terms; (ii) the possibility of selecting a stop list according to the user's criteria; and (iii) the lemmatization of the words contained in the corpus of analysis.

Although the experiment proposed by Barrón and others represents an advance, the results that reach in precision and coverage do not give a relevant balance, in specific his level of precision is of only 0.3, while his coverage reaches in one of its maximum values a 0.8. Given these results, the extracted candidates are not entirely reliable, therefore, is required a revision by an expert.

One of the factors that affects this low performance in the values of precision and recall is the use of a stop list prepared subjectively. In contrast to this criterion, in the experiment reported here was used a stop list extracted from a biomedicine corpus. Such stop list helped to filter good candidates for terms in a phase of selective elimination, p. e: the NP *distant lymph node*, after passing through this phase, is reduced as a *lymph node*, which leads to its consideration as an authentic term candidate. Emphasizing the role of adjectives as terminological units, Acosta & Aguilar (2015), as well as Acosta, Aguilar & Infante (2015), have analyzed their function to describe the attributes associated with a name, giving rise to a categorization of the referent expressed by said name within a specific conceptual domain.

Relational Adjectives Used for Extracting Terms

For this experiment is relevant to have in mind the syntactic and semantic role that relational adjectives play in the construction of biomedical terms. These adjectives, according to Demonte (1999), indicate features that ascribe to the noun within a specific class. Such features can be recognized in NPs such as *muscular inflammation*, *venereal disease* or *psychological disorder*, which follow the sequence: noun + relational adjective.

Taking into account the above, the method that was implemented here considers the use of two annotated corpus with Part-of-Speech (POS) tags, using the **FreeLing** system (Carreras *et al*, 2004). Once such annotation was made, the tags were standardized, in order to regulate the use of the preposition in cases of contraction (in Spanish: *del*), replacing it with the label *PDEL*. Similarly, in most cases, some tags were reduced to two characters, excepting the cases of the verb *estar* (English: *to be*), whose tag is VAE.

After the annotation, a **chunker** was developed to identify NPs that have a more adjective, according to the pattern: <NC> <AQ> *. Now, to eliminate those adjectives that are not relational (specifically, qualifiers), two patterns were implemented that allowed filtering of these:

<RG><AQ>

<VAE><AQ>

where RG, AQ and VAE (according to FreeLing), correspond respectively to adverbs, adjectives and the verb *to be*.

The next step was to apply the relative frequency ratio approach (Manning & Schütze, 2000) between two corpus—one of the biomedical domain and other of general language that serves as a reference—to assign relevance to the words common to both. For this, the formula used was:

$$weight(w_i) = \log_2 \left(1 + \frac{x_{dom}/N_{dom}}{x_{ref}/N_{ref}} \right) \quad (2)$$

where X_{dom} , N_{dom} corresponds to frequency of occurrence of w_i in domain and size of the vocabulary of names and adjectives of the domain, respectively. Similarly, X_{ref} , N_{ref} correspond to the frequency of occurrence of the word w_i and size of the vocabulary of the reference corpus. The unit added to the quotient of said formula is intended to avoid negative values.

On the other hand, the calculation of the relevance of the words that only occur in the domain with frequency of at least 2 was made considering only the logarithm of its frequency:

$$weight(w_i) = \log_2(f_{w_i}) \quad (3)$$

Finally, the words that occurred only in the domain corpus and only once, given that the reference corpus is large (5 million words), the fact that they do not appear in the reference corpus gives scope for considering that the probability of being domain words was high, so the decision to assign them a score of 1:

$$weight(w_i) = \log_2(1+f_{w_i}) \quad (4)$$

For its part, the calculation of *termhood* of each candidate was made through the sum of the relevance of each word present in such candidate: if a NP candidate (in Spanish: *fn*) has a length of n words, $w_1 w_2 \dots w_n$, then the relevance of the candidate is the sum of the weights of all the words w_i :

$$Rank(fn) = \sum_{i=1}^n weight(w_i) \quad (5)$$

The sum of all these values and formulas configures our method of extracting biomedical terms. We apply our method on a corpus with documents collected from **MedLineplus in Spanish** (www.nlm.nih.gov/medlineplus/spanish). The size of the corpus is 1.2 million words. On the other hand, a reference corpus of 5 million words was compiled, which contains news taken from the Mexican newspaper *La*

Jornada (www.jornada.unam.mx), specifically those of the year 2014. For the design of the chunker, it was used the NLTK library (www.nltk.org), developed in Python language (Bird, Klein & Loper, 2009).

Results of the Term Extraction

Without delimiting a specific threshold, and also without eliminating noise in the candidates, in a first phase it was obtained a precision of 33.6% and a recall of 90.4%. For the next phase the noise was removed, thus impacting positively on the results, since a precision of 46.5% and a coverage of 87.4 was achieved. Subsequently, when the C-Value / NC-Value algorithm was implemented to analyze the domain corpus, a precision of 27%, and a recall of 86% was reached.

This method allows establishing different thresholds of relevance (that is, *termhood*) to recover the best candidates. Thus, if a level greater than or equal to 1 is set, a precision of 53.4% and a recall of 85.6% is achieved. The thresholds that are greater than 10 achieve a precision above 70%, with a significant impact on a recall of 16%.

SECOND TASK: EXTRACTION OF HYPONYMS, HYPERNYMS AND MERONYMS

After performing the extraction of terms described in the previous section, the next task that we were going to solve was the identification of hyponyms, hypernyms and meronyms linked to those terms, taking into account the experiments reported by Acosta, Sierra & Aguilar (2015), as well as Acosta & Aguilar (2015, 2016). For carrying out such identification, we reuse our biomedical corpus in Spanish. Likewise, we made modifications to our chunker, in order to recognize NPs composed by nouns, relational adjectives and locative adverbs, considering this class of adverbs can help recognize parts of the human body, e. g. : *the gall bladder is in front of the liver*. The syntactic pattern that represents the noun phrases with a relational adjectives is:

<NC><AQ>

These NP patterns are useful for recognizing simple and complex terms, and also allow deducing possible hyponyms, which are located in a subordinate or specialized category. Previously, it is necessary to remove those NPs with qualifying adjectives, which do not provide candidates for relevant terms. For this, it is considered to elide the following sequences:

<RG><AQ>

<VAE><AQ>

<D.*|P.*|F.*|S.*><AQ><NC>

Using again the tags proposed by **FreeLing**, non-relevant phrase patterns composed of adverbs have been constituted (RG) preceding an adjective (AQ), or the verb *estar* (VAE) introducing an adjective. On the other hand, the <D tags. * | P. * | F. * | S. *> correspond to determinants, pronouns, punctuation marks and prepositions, which are also considered as elements of exception. All these restrictions are useful to reduce the noise that can present false positives at the time of extraction.

Regarding the identification of meronyms, we have considered patterns of nominal phrases that project the axial properties of a reference object (Acosta & Aguilar, 2015; 2016). Under such consideration, both those that include adverbs of place and the preposition of are good candidates, since they refer to specific entities located in a spatial framework represented by the human body. The regular expression to identify this information is:

<RG><PDEL><DA>?<NC><AQ>*

Calculating the Relevance of Identified Candidates

For determining how relevant the candidates we have identified in the analysis corpus are, the relative frequency relationship has been calculated using formula (2). On the other hand, the relevance of the words that only occur in the domain corpus is calculated considering the following formula:

$$weight(fn) = freq(fn) + \left(\sum_{i=1}^n weight(w_i) \right) \quad (6)$$

Where was determined the weight of NPs recognized in such corpus. The classification of candidates for term is done by adding the individual ranges of words present in the identified pattern. In this sense, what is expected is that the higher the frequency of the appearance of noun phrases, the greater their stability as a syntactic structure that projects a biomedical term. It is important to remember that before calibrating the ranges of candidate of NP, these have already been stripped of those non-relevant elements, which facilitates the calculation of their frequency of occurrence.

It should also be added to eliminate the non-relevant vocabulary of the domain corpus, the same criterion used in the term extraction was followed: a contrast was made with a reference corpus previously mentioned, based on news taken from the Mexican newspaper *La Jornada*.

RESULTS AND CONCLUSION

Once the search and collection of the patterns described previously has been carried out, the sets of candidates for hyponyms, hypernyms and meronyms obtained are evaluated. In the case of hyponym/hypernym relationships, the following results were obtained:

Table 1. Precision, recall and f-measure for the detection of hyponyms and hypernyms

Phase	Recall	Precision	F-Measure
Baseline	87%	17%	28%
Chunker	58%	62%	60%
Filter 1	57%	68%	61%
Filter 2, Frequency ≥ 5	42%	80%	55%
Filter 2, Frequency ≥ 6	40%	81%	54%
Filter 2, Frequency ≥ 10	35%	84%	49%
Filter 2, Frequency ≥ 20	27%	90%	42%

What is exposed in this table are the results of precision, recall and F-Measure. The baseline used here consists in extracting candidates to DCs, which introduce PrPs whose heads are verbs such as *to be*, *to characterize*, *to conceive*, *to define* and others similar, according to Sierra *et al* (2008). The proposed chunker managed to achieve a recall of 58% and a precision of 62%. Applying the first causal relationship filter, it is obtained a precision of 68% with a reduction in recall of 57%. Likewise, the application of thresholds of occurrence of hypernyms shows the best results in precision, but as the thresholds increase, the recall is markedly reduced.

In the case of meronym relationships, 5 evaluation phases have been established. In the first, which corresponds to the extraction of specific entities, achieved a precision of 73%. In the next, the precision decreased to 51%. In the remaining phases (3 to 5) they show no changes with respect to their precision. Given these results, the meronyms were extracted with the set obtained from phase 1, in such a way that the candidates from the remaining factions were not analyzed. It should also be added that a general recall of 65% was achieved.

One of the relevant results of this experiment was the level of precision in the recognition of domain terms: as can be seen in table 2, it was higher than 70%. Likewise, it can be observed that the extraction of specific entities tends to converge in phase 5:

Table 2. Levels of precision of candidates for specific entities in meronymy relationships.

Phase	Precision	Concrete Entities	Candidates	Terms
1	73%	308	423	74%
2	51%	566	1109	74%
3	46%	744	1632	72%
4	44%	806	1849	71%
5	43%	821	1891	71%

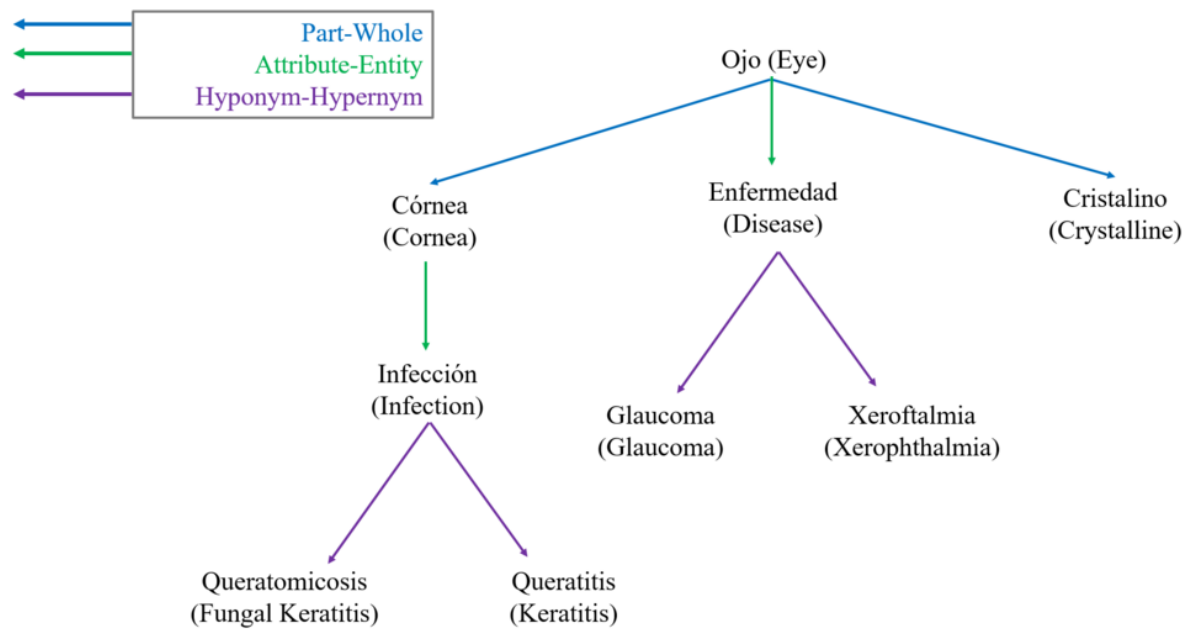
To conclude, it is clear that the mining concepts applied to extract relevant medical information is a complex task that requires many levels of refinements at syntactic and semantic levels, particularly:

- The recognition and filtering of syntactic patterns that show good candidates for terms and definitions, which can be reused when necessary.
- The analysis and inclusion of patterns that indicate other semantic relationships different from those that have been shown here, e. g: causal (*the Helicobacter Pylori is a bacterium that **causes** gastritis and peptic ulcer*), locative (*the pylorus **is situated in** the lower part of the stomach*), telic (*hepatocytes **produce** bile secretions*), and others. Taking these relationships into account will allow the development of a much more detailed model of ontology, thinking in the information obtained through mining previously proposed.
- The application of models such as Markov chains, support vector machines, neural networks and others that allow the implementation of automatic learning methods, in order to provide a relevant degree of independence to the extraction system, which allows it to improve the accuracy of their results, and thereby reduce the human evaluation process.

Therefore, applying the processes of recognition and extraction of DCs, we can obtain terms associated to definitions. Integrating all the lexical-semantic relationships recognized, it is possible to implement an

ontology that allows organizing concepts related to terms. These terms will be extracted from DCs, and the concepts will be derived from analytical definitions. In addition, it is necessary to search synonyms related to these terms. An example of such ontology can be seen in the following figure:

Figure 5. Example of a possible ontology based on medical terms (Aguilar, Sierra & Acosta, 2012)



The Figure 5, we can distinguish a set of sub-terms linked to the main term *ojo* (Eng. *Eye*). These sub-terms operate as nodes, and the possible lexical relations are branches connected with the main term. Thus, based on a lexical relation of Parth-Whole, we can infer that cornea is a constituent of *ojo*. In contrast, the term *enfermedad* (Eng. *Disease*) is an attribute. Finally, the *glaucoma* is a *type of disease* that affects the eyes. This last phase about the generation of a graph that associates the existing relationship between terms is a work still in process.

FUTURE RESEARCH DIRECTIONS

We believe that the results presented here are an empirical support that justifies the implementation of methods for mining concepts on large corpora based on NLP techniques. Although in this case we focus mainly on the identification of terms as well as segments immersed in lexical relations (specifically: hyponyms, hypernyms and meronyms), it should be noted that these lexical units are useful for generating a taxonomy, whose semantic information of the actual use of such units in specialized texts.

Thus, a next step would be to extract the entire DC identified in a text, and infer other semantic relationships contained in it, e. g.: causal relationships such as those of effect-cause (*H. Pylori is a gram-negative bacterium {causes/incites/generates} gastritis*), source-goal (*the gastrointestinal tract starts in the oral cavity and ends in the anus*), containment (*proteins are {formed by/integrated by/contain} amino acids*), and so on. This idea is aligned to the methodology proposed by Buitelaar, Cimiano & Magnini

(2005). Thus, it is possible to employ our processes for recognizing and extracting analytical definitions in DCs in order to build an ontology according the method proposed by Buitelaar, Cimiano, and Magnani.

REFERENCES

- Acosta, O., & Aguilar, C. (2015). Extraction of definitional contexts from restricted domains by measuring synthetic judgements and word relevance. In *Proceedings of the Conference Terminology and Artificial Intelligence TIA 2015* (pp. 161-165). Granada, Spain: Academic Press.
- Acosta, O., Aguilar, C., & Infante, T. (2015). Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus. *Linguamática*, 7(2), 19–34.
- Acosta, O., Sierra, G., & Aguilar, C. (2015). Extracting definitional contexts in Spanish through the identification of hyponymy-hyperonymy relations. In J. Žižka & F. Dařena (Eds.), *Modern Computational Models of Semantic Discovery in Natural Language* (pp. 48–70). IGI Global. doi:10.4018/978-1-4666-8690-8.ch003
- Aguilar, C., & Acosta, O. (2016). Design of an extraction system for definitional contexts from biomedical corpora. In *Proceedings on the First Workshop on Advances in Bioinformatics and Artificial Intelligence: Bridging the Gap*. (pp. 35-41), New York: CEUR Workshop Proceedings.
- Aguilar, C. Acosta, O. Sierra, G. Juárez, S. & Infante, T. (2016). Extracción de contextos definitorios en el área de biomedicine. *Revista de Procesamiento de Lenguaje Natural*, 57, 167-170.
- Ananiadou, S., Kell, D., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12), 571–579. doi:10.1016/j.tibtech.2006.10.002 PMID:17045684
- Ananiadou, S., & McNaught, J. (2006). *Text Mining for Biology and Biomedicine*. Artech House.
- Arp, R., Smith, B., & Spear, A. (2015). *Building ontologies with basic formal ontology*. MIT Press. doi:10.7551/mitpress/9780262527811.001.0001
- Bäck, A. (2000). *Aristotle's Theory of Predication*. Brill. doi:10.1163/9789004321090
- Baker, C., Fillmore, C., & Cronin, B. (2003). The Structure of FrameNet Database. *International Journal of Lexicography*, 16(3), 281–296. doi:10.1093/ijl/16.3.281
- Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. *Proceedings of the COLING-ACL*, 98, 86–90.
- Barrón, A., Sierra, G., Drouin, P., & Ananiadou, S. (2009). An improved automatic term recognition method for Spanish. In *Computational Linguistics and Intelligent Text Processing*, (pp. 125–136). Berlin: Springer.
- Bichindaritz, I., & Akkineni, S. (2006). Concept Mining for Indexing Medical Literature. *Engineering Applications of Artificial Intelligence*, 19(4), 411–417. doi:10.1016/j.engappai.2006.01.009
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing whit Python*. O'Reilly.
- Boukenze, B., Mousannif, H., & Haqiq, A. (2016). Predictive analytics in healthcare system using data mining techniques. *Computer Science and Information Technology*, 1, 1–9.

- Bowers, J. (1993). The Syntax of Predication. *Linguistic Inquiry*, 24(4), 591–636.
- Bowers, J. (2001). Predication. In M. Baltin & C. Collins (Eds.), *The Handbook of Contemporary Syntactic Theory*: 299–333. Blackwell. doi:10.1002/9780470756416.ch10
- Buitelaar, P., Cimiano, P., & Magnini, B. (2005). *Ontology learning from text*. IOS Press.
- Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An open source suite of language analyzers. In *Proceedings of the LREC-2004*. Lisbon, Portugal: ELRA. www.lrec-conf.org/proceedings/lrec2004/
- Chomsky, N. (1981). *Lectures on Government and Binding*. Mouton de Gruyter.
- Cocchiarella, N. (2003). Conceptual realism and the nexus of predication. *Metalogicon*, 16(2), 45–70.
- Demonte, V. (1999). El adjetivo: Clases y usos. La posición del adjetivo en el sintagma nominal. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (Vol. 1, pp. 129–215). Espasa.
- den Dikken, M. (2006). *Relators and Linkers*. MIT Press. doi:10.7551/mitpress/5873.001.0001
- Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multiword terms: The c-value/nc-value method. *International Journal on Digital Libraries*, 3(2), 115–130. doi:10.1007007999900023
- Guarino, N. (1998). Formal Ontology in Information Systems. *Proceedings of FOIS*, 98, 3–15.
- Hoyt, R., Snider, D., Thompson, C., & Mantravadi, S. (2016). IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics. *JMIR Public Health and Surveillance*, 2(2), e157. doi:10.2196/publichealth.5810 PMID:27729304
- Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2), 259–289. doi:10.1075/term.3.2.03kag
- Kaggal, V., Elayavilli, R., Mehrabi, S., Pankratz, J., Sohn, S., Wang, Y., Li, D., Rastegar, M., Murphy, S., Ross, J., Chaudhry, R., Buntrock, J. & Liu, H. (2016). Toward a Learning Health-care System – Knowledge Delivery at the Point of Care Empowered by Big Data and NLP. *Biomedical Informatics Insights*, 8(1), 13–22.
- Kit, Ch., & Liu, X. (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2), 204–229. doi:10.1075/term.14.2.05kit
- Manning, Ch., & Schütze, H. (2000). *Foundations of Statistical Natural Language*. MIT Press.
- Navigli, R., & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2), 151–179. doi:10.1162/089120104323093276
- Poesio, M., & Almuhareb, A. (2008). Extracting Concept Descriptions from the Web: The Importance of Attributes and Values. In P. Buitelaar & Ph. Cimiano (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 29–44). IOS Press.
- Rothstein, S. (1983). *The Syntax Forms of Predication* (Tesis de Doctorado). Cambridge, MA: MIT.
- Sainani, K. (2008). Mining biomedical literature: Using computers to extract knowledge nuggets. *Biomedical Computation Review*, (July), 16–27.

- Sierra, G. (2009). Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, 1(2), 13–37.
- Sierra, G. Alarcón, R. & Aguilar, C. (2006). Extracción automática de contextos definitorios en textos especializados. *Revista de Procesamiento de Lenguaje Natural*, 37, 351–352.
- Sierra, G., Alarcón, R., Aguilar, C., & Bach, C. (2008). Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1), 74–98. doi:10.1075/term.14.1.05sie
- Smith, B. (2004). Beyond Concepts: Ontology as Reality Representation. *Proceedings of International Conference on Formal Ontology and Information Systems FOIS 2004*.
- Velardi, P., Faralli, S., & Navigli, R. (2013). OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3), 665–707. doi:10.1162/COLI_a_00146
- Zweigenbaum, P., Demner-Fushman, D., Yuand, H., & Cohen, K. (2007). Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*, 8(5), 358–375. doi:10.1093/bib/bbm045 PMID:17977867

ADDITIONAL READING

- Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*: 57–64. Stroudsburg, PA: ACL Publications. 10.3115/1034678.1034697
- Berry, M., & Kogan, J. (2010). *Text mining. Applications and theory*. London, UK. Wiley.
- Feldman, R. & Sanger, J. (2007). *The text mining handbook*. Cambridge, UK. Cambridge University Press.
- Hearst, M. (1999). Untangling Text Data Mining. In *Proceedings of the 37th Annual Meeting of the ACL on Computational Linguistics ACL'99*. (pp. 3–10). Stroudsburg, PA, USA. ACL Publications.
- Bodenreider, O. (2001). *Medical ontology research. A report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications*, NLM, USA: <https://pdfs.semanticscholar.org/51e4/8185247355536bd9420a9d5005c9e50d5774.pdf>
- Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.
- Faure, D., & Nedellec, C. (1998). A Corpus-based Conceptual Clustering Method for Verb Frames and Ontologies. In *Proceedings of the LREC Workshop on adapting lexical and corpus resources to sublanguages and applications*. (pp. 5–12). Granada, Spain: www.elra.info/en/lrec/proceedings
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press. doi:10.7551/mitpress/7287.001.0001
- Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic discovery of part–whole relations. *Computational Linguistics*, 32(1), 83–135.
- Gruber, T. (1993). Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 199–220. doi:10.1006/knac.1993.1008

Hearts, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. (pp. 539-545). Nantes, France.

Hirst, G. (2009). Ontology and the lexicon. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 269–292). Springer. doi:10.1007/978-3-540-92673-3_12

Huang, Ch., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., & Prévost, L. (2010). *Ontology and the lexicon. A natural language processing perspective*. Cambridge University Press. doi:10.1017/CBO9780511676536

Maynard, D., Li, Y., & Peters, W. (2008). NLP Techniques for Term Extraction and Ontology Population. In P. Buitelaar & Ph. Cimiano (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 107–127). IOS Press.

Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. MIT Press.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of Conference on Computational Linguistics*. (pp. 113-120). Stroudsburg, PA, USA. ACL Publications. 10.3115/1220175.1220190

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–48). LEA Publications.

Smith, B. (2003). Ontology. In L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and Information* (pp. 155–166). Blackwell.

Snow, R., Jurafsky, D., & Ng, A. (2006). Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*. (pp. 801–808). Stroudsburg, PA, USA. ACL Publications.

Tan, A. 1999. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD Workshop on Knowledge discovery from Advanced Databases (KDAD'99)*. (pp. 71-76). Beijing, China.

Vossen, P., & Copestake, A. (1993). Defaults in lexical representation. In T. Briscoe, V. Paiva, & A. Copestake (Eds.), *Inheritance, Defaults and the Lexicon* (pp. 246–274). Cambridge University Press.

Wilks, Y. Sator, B. & Guthrie, L. (1996). *Electric Words*, Cambridge, Mass: MIT Press.

Winston, M., Chaffin, R., & Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4), 417–444. doi:10.1207/15516709cog1104_2

KEY TERMS AND DEFINITIONS

Concept: A mental construction which contains information about a specific entity based on our knowledge and our experience.

Concept Mining: A kind of task derived from the knowledge discovery, focused to recognize patterns in large collections of unstructured data (that is, linguistic corpora), which contain relevant information associated to concepts. For solving such task, it is employed different methods based on NLP (e.g., tokenization, lemmatization, POS tagging, syntactic and semantic analysis, and others).

Definitional Context: A discursive fragment constituted by a term, a definition, and linguistic or metalinguistic connector as predicative phrases, typographical markers, or pragmatic patterns.

Hyponymy-Hypernymy: A kind of hierarchical lexical relation where is established a conceptual subordination between the meaning of two or more words (e.g., **a robin** [hyponym] *is a kind of bird* [hypernym]).

Lexical Relationship: A conceptual link established between words grouped around a lexicon.

Meronymy: A semantic relation specific to linguistics, distinct from the similar metonymy. A meronym denotes a constituent part of, or a member of something (e.g., **wheels** *are part of an automobile*; **a policeman** *is a part of a police department*, etc.).

Ontology: A system of representation that shows the formal naming and definitions of the categories associated to entities. It describes the properties and relations between the concepts refer by such entities, situated in a specific or general knowledge domain.

Term: A word or compound words that designates a specific entity, event or relation into the framework of a scientific or technical knowledge domain.

Text Mining: A process of deriving high-quality information from text. This kind of task involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Section 7

Digital and Wireless Communications

Overview of Machine Translation Development

Irene Rivera-Trigueros

 <https://orcid.org/0000-0003-4877-4083>

Department of Translation and Interpreting, University of Granada, Spain

María-Dolores Olvera-Lobo

 <https://orcid.org/0000-0002-0489-7674>

Department of Information and Communication Sciences, University of Granada, Spain

Juncal Gutiérrez-Artacho

 <https://orcid.org/0000-0002-0275-600X>

Department of Translation and Interpreting, University of Granada, Spain

INTRODUCTION

Information is increasingly global, which brings with it the growth in a non-English speaking public and, as such, a demand for tools that allow users to access this information. Faced with this scenario, the development of new tools and technologies is a key element. This is the case of Machine Translation (MT), one of the most popular and used tools to overcome linguistic barriers. As a result, linguistic companies are increasingly offering services related to MT as, in many cases, it is impossible to meet the demand for translations only with human translators. In addition, MT systems can be integrated with other tools and resources such as Translation Memories—databases which store already translated text for future use—or multilingual terminology databases—also known as glossaries—in order to improve the efficacy of the translation process.

In this chapter an overview of MT development since its early beginnings in the late 1950s until the present day is proposed. The evolution of MT techniques will be reviewed starting from the traditional approaches—Statistical and Rule-Based MT—and finishing with the latest trends, which include hybrid MT systems, Adaptive MT and Neural MT.

BACKGROUND

Brief History of Machine Translation

The origins of Machine Translation date back to the 17th century, as during this time an interest for creating an universal language based on logic principles and iconic symbols and for developing mechanical dictionaries arises (Hutchins, 2007). During the following centuries, there were some proposals to create international artificial languages, among which stood out Esperanto (Hutchins, 1995; Hernández-Mercedes, 2002). However, there were no real attempts for automating translation until 1933, when two patents were issued for machines functioning as multilingual mechanical dictionaries, designed to find

DOI: 10.4018/978-1-7998-3479-3.ch060

the equivalent of a word in another language. One patent was issued in France, to Georges Artsouni, and the other one in Russia, to Petr Trojanskij. During the 1940s Andrew Booth and Warren Weaver developed the first ideas which involved computers for translating natural languages and worked with Richard H. Richens on morphological analysis for a mechanical dictionary (Hutchins, 1995).

In 1951 Yehoshua Bar-Hillel, from the Massachusetts Institute of Technology (MIT), outlined some of the basic approaches to MT in a state-of-the-art-report and one year later the first MT conference was held at MIT. The conference evidenced that good quality MT was not possible, as long-term research was needed. Therefore, at that stage, human intervention was required before (pre-editing) or after (post-editing) the process. In addition, during the conference it was also highlighted that a public demonstration was necessary in order to attract funding (Hutchins, 2007; Hernández-Mercedes, 2002). Two years later, in 1954, the first public demonstration of an MT system was carried out as a result of a joint project between IBM and Georgetown University. Despite not having a remarkable scientific value, this demonstration was the starting point for large-scale funded MT research in the USA and for the developing of MT projects around the world, especially in the USSR (Hutchins, 1995).

During the following ten years, there were two main approaches regarding MT research. On the one hand, empirical approaches focused on statistical methods in order to discover grammatical and lexical regularities. On the other hand, theoretical approaches were based on linguistic research, which led to the beginnings of computational linguistics (Hutchins, 2007).

Even though during this decade research on MT field seemed promising, the linguistic and semantic problems became more and more noticeable. Therefore, the National Science Foundation was asked by government sponsors in the USA to assess future prospects of MT. For this purpose, the Automatic Language Processing Advisory Committee (ALPAC) was formed in 1964. The ALPAC report in 1966 pointed out that MT was slower, less accurate and twice as expensive as human translation. Consequently, the ALPAC report meant a significant reduction in MT research and development (Dorr, Jordan, & Benoit, 1999). Despite the impact of the ALPAC report, some research groups at the University of Texas and at Wayne State University continued researching on MT in the USA, which were focused on translating scientific and technical materials from Russian into English (Hutchins, 2007). However, between 1967 and 1977, MT research was carried out mainly in Canada and Europe. There was a high demand for English-French translations in Canada due to its bicultural policy. For its part, in the European Community, it was necessary to translate scientific, technical, administrative and legal documentation into all the languages of the European Community. During the decade of 1970, at Montreal, researches of the TAUM project (Traduction Automatique de l'Université de Montréal) designed Méteo, a system for translating weather forecast (Dorr et al., 1999). During the same years, the European Community purchased an MT system called Systran—developed by Peter Tomas—and, at the end of the 1970s, decided to establish a research project called EUROTRA, whose aim was to create an MT system for all the official languages of the European Community (Alcina Caudet, 2011).

In the 1980s, Systran continued to evolve and include more languages while new systems such as LOGOS or METAL entered the market. Special-purpose MT systems—designed for particular environments—were also developed. That was the case of SPANAM and ENGSPAN, two systems developed by the Pan American Health Organization (Hutchins, 2007).

During the 1980s, research groups continued working on MT. For instance, the GETA (Groupe d'Études pour la Traduction Automatique) group developed the Ariane system, which influenced MT projects around the world. In 1986, the research prototype of the Mu system, which was conceived at the University of Kyoto, became an operational system used by the Japanese Information Center for Science and Technology. This project influenced Japanese computer companies, which started to invest

and develop computer-aided translation software (Hutchins, 2007). The increase of the availability of personal computers and text processors resulted in cheaper MT systems, which were exploited by companies such as ALPS, Globalink, NEC and Sanyo, among others (Hernández-Mercedes, 2002). During this years, research on MT continued and projects were developed in Europe, Korea, Taiwan, China, North America and the Soviet Union (Hutchins, 1995).

In the 1990s, the use of MT systems by commercial agencies, government services and multinational companies was consolidated and Systran, Logos, METAL and ATLAS systems dominated the market. In addition, new fields of research, such as software and website localization, emerged and, given the need of immediate translations, MT—along with other resources like Translation Memories—become a popular tool in this areas. Numerous systems for personal computers appeared and started to be commercialized, normally in three basic versions: for large corporations ('enterprise' systems), for independent translators ('professional' systems) and for non-translators ('home use') (Hutchins, 2007). In addition, from 1995 on, internet-based translation services appeared and, even though the quality was generally poor, they met the demand for immediate translation for information purposes (Hutchins, 2007).

From the 2000s on, MT has been a very prolific research field (España-Bonet & Costa-jussà, 2016). As a result, new approaches continuously emerge and different technologies and techniques are being quickly developed in order to improve MT systems as much as possible. In the following section, MT strategies and techniques will be revised starting from the first traditional approaches and finishing with the newest tendencies in this field.

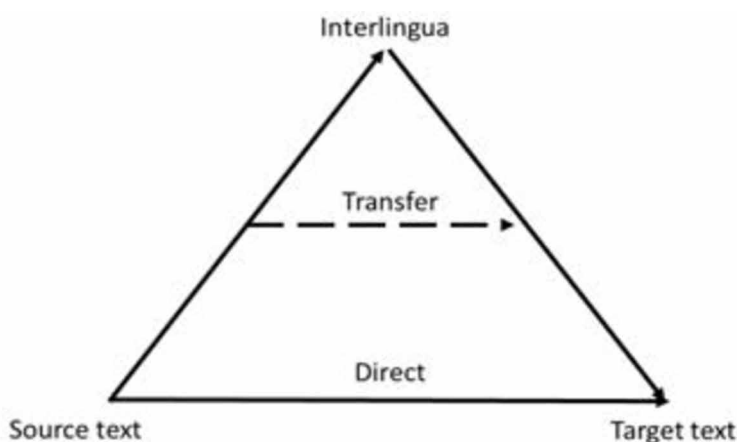
MACHINE TRANSLATION APPROACHES

Traditionally, two main basic approaches were distinguished concerning MT: the rule-based approach and the corpus-based approach. However, in the recent past years MT research has explored new techniques and architectures, resulting in new approaches such as Hybrid MT, Adaptive MT and Neural MT.

Rule-Based Machine Translation

Rule-Based Machine Translation (RBMT) systems use monolingual and bilingual dictionaries and grammar rules as the basis to generate translations (Hutchins & Somers, 1992; Okpor, 2014). Three basic MT strategies can be identified with regard to RBMT systems: direct, interlingual and transfer. These approaches are traditionally represented using the Vauquois triangle (Figure 1).

Figure 1. Vauquois triangle (adapted from Vauquois, 1968)



Direct Approach

The direct approach was used by the first generation of MT systems (Hutchins & Somers, 1992). There are not intermediate stages in the translation process, that is to say, the source language input is replaced word by word in the target language. Some local reordering rules are followed in order to achieve a more precise target language output (Hutchins & Somers, 1992; Prószéky, 2005).

The first MT system—called IBM701 and developed jointly by Georgetown University and IBM—was based on the direct approach (Peng, 2013).

Interlingual Approach

The interlingual approach analyses the source text using an intermediate representation, which includes all the necessary information for generating the target text. This neutral and abstract representation acts as a projection from the source text which, at the same time, is the basis for the generation of the target text (Hutchins & Somers, 1992; Okpor, 2014).

The UNITRAN system, developed at the Artificial Intelligence Laboratory of the MIT (Dorr, 1987), and the KANT system, developed at Carnegie Mellon University (Nyberg, Kamprath, & Mitamura, 1998) are examples of MT systems using interlingual architecture.

Transfer Approach

The transfer approach transforms the source language into an intermediate representation and an equivalent representation is generated for the target languages using dictionaries and grammars (Hutchins & Somers, 1992; Okpor, 2014; Tripathi & Sarkhel, 2010). The difference with interlingual MT is that the transfer approach depends on the language pair involved (Okpor, 2014). The translations are carried out in three stages: analysis, transfer and generation. In the first stage the source text is analyzed based on linguistic information in order to produce its representation; the next stage aims to transfer the source language representation into the semantic/syntactic structure of the target language. In the final stage a target language morphological analyzer is used to generate the final output (Okpor, 2014; Tripathi & Sarkhel, 2010).

The MANTRA system, developed by the Applied Artificial Intelligence Group (AAI) of C-DAC (Center for Development of Advanced Computing) in Pune (India), translates documents from a variety of domains from English to Hindi using a transfer-based approach (Gehlot et al., 2015). The Dan2eng system, for Danish-English, also uses a transfer-based approach to translate unrestricted text from a wide range of genres (Bick, 2007).

Corpus-Based Machine Translation

From 1989, new methods and strategies—known as corpus-based methods—emerged and gained popularity, replacing the dominant RBMT systems (Hutchins, 2007). Corpus-Based Machine Translation (CBMT) systems use bilingual parallel corpora, consisting in texts paired with their translations, which are used to obtain translation knowledge with the aim to achieve the highest possible level of accuracy (Ballabh & Chandra Jaiswal, 2015; Okpor, 2014). We can distinguish between two main approaches concerning CBMT: Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT), although it is important to mention that Somers (1998) argues that SMT is example-based

approach and Hutchins (2007) states that both approaches converge in many respects and, consequently, isolating the distinctive features of both approaches is a very complex task.

Statistical Machine Translation

SMT is based on machine learning, a learning algorithm is applied to large volume of corpora—which include many samples of human translations—so that the system is able to learn automatically how to translate without human intervention (Dajun & Yun, 2015; Lopez, 2008). Therefore, SMT is based on probabilistic models that take into account both the relations in the source and the target language to generate the output (España-Bonet & Costa-jussà, 2016). Depending on the minimal unit for translation, three different approaches can be distinguished: word-based, phrase based or syntax-based (España-Bonet & Costa-jussà, 2016; Tripathi & Sarkhel, 2010).

There are many examples of SMT systems. For instance, the open-source SMT Moses allows to train translation models for any language pair (Koehn et al., 2007). Within the commercial MT systems, the company SDL International acquired Language Weaver, an SMT system developed at the University of Southern California's Information Sciences Institute, which can be integrated with SDL computer-assisted translation tools and it is available for 75 language pairs (Alcina Caudet, 2011; Benjamin, Gerber, Knight, & Marcu, 2003). The well-known companies Microsoft and Google used the statistical approach in their MT systems—Microsoft Translator and Google Translate. However, since 2016 both companies shifted towards artificial intelligence and both MT systems were updated to implement Neural Machine Translation (NMT) (Microsoft blog editor, 2018; Wu et al., 2016).

Example-Based Machine Translation

EBMT is based on translation by analogy. An EBMT system extracts from a corpus or database segments of texts written in the source language in order to generate a text in the target language with the same meaning (Hutchins, 2005; Okpor, 2014; Somers, 1999). EBMT entails matching new input—sequences of words—against the database so that suitable examples are extracted and recombined to generate the correct translation in the target language (Okpor, 2014; Somers, 1999).

As previously stated, it is difficult to clearly distinguish between SMT and EBMT systems and, as a consequence, EBMT systems and SMT normally share features of their paradigms. This is the case of Cunei, a hybrid EBMT and SMT system (Phillips, 2011) and CBU-EBMT, a complete end-to-end system available under the GNU General Public License (Brown, 2011).

Hybrid Machine Translation

Hybrid Machine Translation (HMT) combines system components of RBMT and SMT into new systems with the aim to join the advantages of two approaches (Hunsicker, Yu, & Federmann, 2012; Tambouratzis, 2014; Thurmair, 2009). In this sense, HMT systems try to overcome the problems caused by these two approaches separately in order to increase the quality and accuracy of translations (Gutiérrez-Artacho, Olvera-Lobo, & Rivera-Trigueros, 2018; Hunsicker et al., 2012; Tambouratzis, 2014).

Hybridization of MT can be carried out following several techniques but, in general terms, we can distinguish between hybridization led by an RBMT system or an SMT system (Costa-Jussà & Fonollosa, 2015; España-Bonet & Costa-jussà, 2016). On the one hand, SMT systems can be integrated with RBMT systems in order to improve syntax and grammar. This integration can be performed at a superficial

level—by using rules at pre/post-processing—or at a deeper level—integrating RMBT features into the core translation model of an SMT system (España-Bonet & Costa-jussà, 2016). On the other hand, RBMT systems can integrate data information to improve lexical selection and fluency (Costa-Jussà & Fonollosa, 2015; España-Bonet & Costa-jussà, 2016). Three classes that summarize the integration of both approaches can be distinguished: enriching the basic resources of RBMT with data obtained from corpora; including data-based modules in the RBMT process and using RBMT as a backbone, which consist of the combination of individual systems where the rule-based one dominates (España-Bonet & Costa-jussà, 2016).

In 2009, AppTek launched the first complete HMT system, called TranSphere, which provided a full integration of RBMT and SMT methodologies (Boretz, 2009). The same year, Systran released SYS-TRAN Enterprise Server 7, the first HMT server of the company combining the RBMT and the SMT approaches (Moran & Donlan, 2009). Nowadays, both companies have implemented the neural approach into their MT systems. In 2012, ProMT presented the ProMT DeepHybrid system, which was based on a RBTM engine which incorporates an statistical phrase-based post-editing component (Molchanov, 2012).

Adaptive Machine Translation

The idea behind Adaptive MT is that the system learns from user feedback in order to avoid repeating the same mistakes in future sentences (Farajian, Bertoldi, & Federico, 2014). Therefore, this paradigm is focused on the post-editing phase, as the systems incorporate the human feedback immediately via online updates (Denkowski, Lavie, Lacruz, & Dyer, 2014).

SDL released SDL Trados Studio 2017, which is designed to be used when no translation memory match is available. The MT engine is updated as users accept new suggestions—in a similar way as a translation memory will be updated—and MT mistakes do not need to be corrected every time, which reduces the post-editing effort. SDL claims that this technology shift means MT and translation memory technology start converging (Ferguson, 2016; Whale, 2017).

Neural Machine Translation

Neural Machine Translation (NMT) is a new approach that aims to train a large neural network that reads a sentence and generates a correct translation (Bahdanau, Cho, & Bengio, 2015). NMT is based on the encoder-decoder model. The encoder reads a source sentence and extracts a fixed-length representation and the decoder generates a correct translation from the previous encoded vector (Bahdanau et al., 2015; Cho et al., 2014). In addition, neural networks can be trained for specializing in specific domains (Wołk & Marasek, 2015).

As previously mentioned, some of the main MT companies have shifted towards NMT. Google launched in 2016 Google's Neural Machine Translation system (GNMT), whose network includes 8 encoders and 8 decoders layers (Wu et al., 2016). Recently, Google is working on hybrid models combining different NMT architectures to develop what they have called the RNMT+ model (Chen et al., 2018). On the other hand, Systran introduced Pure Neural Machine Translation (PNMT) in the same year, which is now available in Systran Enterprise Server and it also embedded in mobile applications (Crego et al., 2016; Systran, 2017). In addition, Systran share research results in OpenNMT, an open source initiative for neural machine translation and neural sequence modeling (Klein, Kim, Deng, Senellart, & Rush, 2017). Microsoft Translator adopted the NMT approach in 2016, which was also implemented by Skype Translator and the Microsoft Translator app for mobile devices (Microsoft Translator, 2016). In May

2018, Microsoft updated their MT system for allowing users to customize the NMT systems (Microsoft blog editor, 2018). Finally, other well-known companies such as Facebook and Amazon are also using NMT in their translation services (Lample, Ott, Conneau, Denoyer, & Ranzato, 2018; Lu et al., 2018).

Summary of MT systems

There are numerous MT systems available, either open-source systems or commercial systems. The European Association of Machine Translation (EAMT) offers a comprehensive directory of commercial MT software on its website (EAMT, 2010). In the previous sections some MT systems have been referred to illustrate the approaches described. Table 1 shows a brief summary of the different MT systems mentioned in the previous sections.

Table 1. MT systems summary

MT Approaches		MT systems	Available at:
RBMT	Direct approach	IBM701	Not available. See IBM press release (January 8, 1954) for more information.
	Interlingual approach	UNITRAN	Not available. See Dorr (1987) for more information.
		KANT	KANT project website
	Transfer approach	MANTRA	CDAC website
		Dan2eng	SDU website
CBMT	SMT	Moses	Moses website
		Language Weaver	SDL Language Weaver website
	EBMT	Cunei	Cunei website
		CBU-EBMT	Not available. See Brown (2011) for more information
HMT		TranSphere	TranSphere website
		ProMT	ProMT website
Adaptive MT		SDL Trados Studio 2017	SDL website
Neural MT		Google Translate	Google Translate online translator
		Systran	Systranet online translator
		Microsoft Translator	Microsoft Bing online translator

SOLUTIONS AND RECOMMENDATIONS

There is no doubt that many advances have been made concerning MT. New technologies and systems are rapidly emerging and evolving achieving significant milestones. However, MT has still its limitations. For instance, NMT is computationally expensive as its training configurations often takes a few weeks (Crego et al., 2016). In addition, NMT can make errors such as removing words, mistranslating proper names or rare terms or not taking into account the context of the translation (Wu et al., 2016). As a consequence, measuring the performance of MT system will be required to be able to progress in its research, development and improvement. To this end, Shaw and Gros (2007) list in their work a number of methods—including human and automatic criteria—that can be used to assess the efficacy of MT. In addition, as MT, in general terms, does not reach the quality of human translations post-editing has

become essential part of the translation process if we wish to obtain translations of the highest possible quality (Gutiérrez-Artacho et al., 2018). In this sense, developing post-editing guidelines will help to increase the efficacy of this process. Several authors offers post-editing guidelines (Hu & Cadwell, 2016; Mesa-Lao, 2013; O' Brien, 2009; TAUS, 2010; Wagner, 1985), although further research, specifically for MT in specific domains is still needed.

FUTURE RESEARCH DIRECTIONS

It is clear that MT is an evolving field and new approaches will rapidly emerge. Nowadays, the main research areas are focused on NMT. However, there are new works that aims to integrate NMT technology into interactive machine translation framework to foster human-computer collaboration (Peris et al., 2017). In fact, there is already one available system—Lilt—that uses the artificial intelligence from NMT and the real-time feedback of Adaptive MT to increase translator's productivity (Lilt, 2017). It can be expected that in the near future ever more companies will research on the joint implementation of this two recent approaches giving birth to a new generation of hybrid architectures. To finish with, in relation to MT evaluation, nowadays, the most used metrics are targeted to SMT systems (Way, 2018). Therefore, as new approaches arise, evaluation metrics adapted to the new features of MT systems are expected to be developed as well. Moreover, new research areas are focused on integrating MT with web localization and the emerging field of transcreation—translating and adapting a message from a source language to evoke the same emotional response in the target language— in order to assess the impact of MT with regards to the post-editing process.

CONCLUSION

MT is clearly an active and growing field of research, long gone are the days of the ALPAC report when there were no feasible prospects for MT. Global access to information and the increase of the non-English speaking audience imply the demand for tools allowing users to secure information from a wide range of resources. MT has proven its potential for overcoming linguistic barriers for professional translators, companies, institutions and private users. As a consequence, the demand for MT is generalized and growing, especially in the multilingual context of the Internet.

MT does not limit itself to the linguistic field, as it could turn really useful in fields such as Medicine or Information Retrieval, among others. For instance, MT results especially convenient when there is a high demand for translating large amounts of texts and human translators alone could not satisfy it due to time or cost constraints. In addition, MT systems, can be integrated together with other resources such as Translation Memories, glossaries or terminology databases for contributing to carry out the translation process in a quicker and more efficient manner. In the future, we can expect an increasing demand for MT services, which will lead to new advances and technologies. However, the quality of machine translations is still deficient, compared to professional human translations. As a consequence, new evaluation metrics and methodologies should be developed to assess MT quality, taking into account human and automatic criteria. As nowadays post-editing is still required to achieve high quality translations, there should also be a focus of research on this field—including competences, guidelines and the role of post-editors within the translation process.

REFERENCES

- Alcina Caudet, M. A. (2011). Los traductores automáticos en la red. In *El español en el mundo. Anuario del Instituto Cervantes 2010-2011*. Instituto Cervantes. Retrieved from https://cvc.cervantes.es/lengua/anuario/anuario_10-11/alcina/p01.htm
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*. Retrieved from <https://arxiv.org/pdf/1409.0473.pdf>
- Ballabh, A., & Chandra Jaiswal, U. (2015). A study of machine translation methods and their challenges. *International Journal of Advance Research In Science And Engineering*, (4).
- Benjamin, B., Gerber, L., Knight, K., & Marcu, D. (2003). Language Weaver : The Next Generation of Machine Translation. Retrieved from <http://www.mt-archive.info/MTS-2003-Benjamin.pdf>
- Bick, E. (2007). Dan2eng: Wide-Coverage Danish-English Machine Translation. In *Proceedings of Machine Translation Summit XI*. Academic Press.
- Boretz, A. (2009). SpeechTechMag.com: AppTek Launches Hybrid Machine Translation Software. Retrieved August 20, 2018, from <https://www.speechtechmag.com/Articles/News/Speech-Technology-News-Features/AppTek-Launches-Hybrid-Machine-Translation-Software-52871.aspx>
- Brown, R. D. (2011). The CMU-EBMT machine translation system. *Machine Translation*, 25(2), 179–195. doi:10.1007/s10590-011-9095-8
- Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., & Foster, G. ... Hughes, M. (2018). The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. Retrieved from <https://arxiv.org/abs/1804.09849>
- Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*. Retrieved from <https://arxiv.org/abs/1409.1259>
- Costa-Jussà, M. R., & Fonollosa, J. A. R. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1), 3–10. doi:10.1016/j.csl.2014.11.001
- Crego, J. M., Kim, J., Klein, G., Rebollo, A., Yang, K., & Senellart, J. ... Zoldan, P. (2016). SYSTRAN's Pure Neural Machine Translation Systems. Retrieved from <https://arxiv.org/abs/1610.05540>
- Dajun, Z., & Yun, W. (2015). Corpus-based Machine Translation : Its Current Development and Perspectives. *International Forum of Teaching and Studies*, 11(1), 90–95.
- Denkowski, M., Lavie, A., Lacruz, I., & Dyer, C. (2014). Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter. Association for Computational Linguistics. Retrieved from <https://github.com/redpony/cdec/tree/>
- Dorr, B. (1987). UNITRAN: An Interlingual Approach to Machine Translation. *AAAI-87 Proceedings*, 534–539.
- Dorr, B. J., Jordan, P. W., & Benoit, J. W. (1999). A Survey of Current Paradigms in Machine Translation. *Advances in Computers*, 49(C), 1–68. doi:10.1016/S0065-2458(08)60282-X

EAMT. (2010). European Association for Machine Translation. Compendium of Translation Software. Retrieved from <http://eamt.org/compendium/index.php>

España-Bonet, C., & Costa-jussà, M. R. (2016). Hybrid Machine Translation Overview. In M. Costa-jussà, R. Rapp, P. Lambert, K. Eberle, R. Banchs, & B. Babych (Eds.), *Hybrid Approaches to Machine Translation. Theory and Applications of Natural Language Processing* (pp. 1–24). Springer., doi:10.1007/978-3-319-21311-8_1.

Farajian, M. A., Bertoldi, N., & Federico, M. (2014). Online Word Alignment for Online Adaptive Machine Translation. In *EACL-2014: 14th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Humans and Computer-assisted Translation* (pp. 84–92). Association for Computational Linguistics., doi:10.3115/v1/W14-0313.

Ferguson, N. (2016). AdaptiveMT for SDL Trados Studio 2017: a self-learning machine translation engine - SDL Trados Blog. Retrieved August 20, 2018, from <https://blog.sdltrados.com/adaptivemt-sdl-trados-studio-2017-transformational-mt-technology/>

Gehlot, A., Sharma, V., Singh, S., Kumar, A., Vidyapith, B., & Rajasthan, I. (2015). Hindi to English Transfer Based Machine Translation System. *International Journal of Advanced Computer Research*, 5(19).

Gutiérrez-Artacho, J., Olvera-Lobo, M.-D., & Rivera-Trigueros, I. (2018). *Human Post-editing in Hybrid Machine Translation Systems: Automatic and Manual Analysis and Evaluation*. Springer., doi:10.1007/978-3-319-77703-0_26

Hernández-Mercedes, P. (2002). En torno a la traducción automática. *Cervantes*, 2, ●●●. https://www.infoamerica.org/documentos_pdf/troyanskii4.pdf

Hu, K., & Cadwell, P. (2016). A Comparative Study of Post-editing Guidelines. *Baltic Journal of Modern Computing*, 4(2), 346–353.

Hunsicker, S., Yu, C., & Federmann, C. (2012). Machine Learning for Hybrid Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 312–316.

Hutchins, J. (1995). Machine Translation: A brief History. In E. F. K. Koerner & R. E. Asher (Eds.), *Concise history of the language sciences: from the Sumerians to the cognitivists* (pp. 431–445). Pergamon Press., doi:10.1016/B978-0-08-042580-1.50066-0.

Hutchins, J. (2005). Example-based machine translation: a review and commentary. doi:10.1007/10590-006-9003-9

Hutchins, J. (2007). Machine translation : A concise history. *Machine Translation*, 13(1 & 2), 1–21.

Hutchins, W. J., & Somers, H. L. (1992). An Introduction to Machine Translation. *The Library*, 57(1), 377. doi:10.2307/3721978

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. Retrieved from <https://arxiv.org/abs/1701.02810>

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., & Bertoldi, N. ... Cornell, E. H. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. ITC-irst 2. Retrieved from <http://www.statmt.org/moses/>

Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-Based & Neural Unsupervised Machine Translation. Retrieved from <https://arxiv.org/abs/1804.07755>

Lilt. (2017). Lilt Launches First-Ever Adaptive Neural Machine Translation System | Slator. Retrieved August 20, 2018, from <https://slator.com/press-releases/lilt-launches-first-ever-adaptive-neural-machine-translation-system/>

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(8), 49. doi:10.1145/1380584.1380586

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., & Sun, J. (2018). A neural interlingua for multilingual machine translation. Retrieved from <https://arxiv.org/abs/1804.08198>

Mesa-Lao, B. (2013). Introduction to post-editing – The CasMaCat GUI 1 . Introduction : Why post-editing MT outputs ? Eamt 2008.

Microsoft Blog Editor. (2018). Customized neural machine translation with Microsoft Translator - Microsoft Research. Retrieved August 17, 2018, from <https://www.microsoft.com/en-us/research/blog/customized-neural-machine-translation-microsoft-translator/>

Microsoft Translator. (2016). Microsoft Translator launching Neural Network based translations for all its speech languages – Translator. Retrieved August 20, 2018, from <https://blogs.msdn.microsoft.com/translation/2016/11/15/microsoft-translator-launching-neural-network-based-translations-for-all-its-speech-languages/>

Molchanov, A. (2012). PROMT DeepHybrid system for WMT12 shared translation task. In *Proceedings of the 7th Workshop on Statistical Machine Translation* (pp. 345–348). Montréal: Association for Computational Linguistics.

Moran, J., & Donlan, B. (2009). SYSTRAN Introduces hybrid machine translation solution for enterprises | SYSTRAN – Translation Technologies. Retrieved August 20, 2018, from <https://www.systransoft.com/systran/news-and-events/press-release/hybrid-machine-translation-solution-for-enterprises/>

Nyberg, E., Kamprath, C., & Mitamura, T. (1998). The KANT Translation System. Localization Industry Standards Association (LISA) Newsletter, 2. Retrieved from <http://www.lti.cs.cmu.edu/Research/KANT>

O’ Brien, S. (2009). Researching and Teaching Post-Editing. Retrieved from <http://www.mt-archive.info/MTS-2009-OBrien-ppt.pdf>

Okpor, M. D. (2014). Machine Translation Approaches: Issues and Challenges. *IJCSI International Journal of Computer Science Issues*, 11(5), 159–165.

Peng, L. (2013). *A Survey of Machine Translation Methods* (Vol. 11). TELKOMNIKA.

Peris, A., Domingo, M., & Casacuberta, F. (2017). Interactive neural machine translation. *Computer Speech & Language*, 45, 201–220. doi:10.1016/j.csl.2016.12.003

Phillips, A. B. (2011). Cunei: Open-source machine translation with relevance-based models of each translation instance. *Machine Translation*, 25(2), 161–177. doi:10.1007/s10590-011-9109-6

Prószeréky, G. (2005). Machine Translation and the Rule-to-Rule Hypothesis. In K. Károly & Á. Fóris (Eds.), *New Trends in Translation Studies (In Honour of Kinga Klaudy)* (pp. 207–218). Akadémiai Kiadó.

- Shaw, F., & Gros, X. (2007). Survey of Machine Translation Evaluation. Survey of Machine Translation Evaluation. Retrieved from http://www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf
- Somers, H. (1999). Review Article: Example-based Machine Translation. *Machine Translation*, 14(2), 113–157. doi:10.1023/A:1008109312730
- Somers, H. L. (1998). “New paradigms” in MT: the state of play now that the dust has settled. 10th European Summer School in Logic, Language and Information, Workshop on Machine Translation.
- Systran. (2017). Understand Neural Machine Translation | SYSTRAN – Translation Technologies. Retrieved August 20, 2018, from <https://www.systransoft.com/understand-neural-machine-translation/>
- Tambouratzis, G. (2014). Comparing CRF and template-matching in phrasing tasks within a Hybrid MT system. Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra), 7–14. doi:10.3115/v1/W14-1003
- TAUS. (2010). MT Post-editing Guidelines. Retrieved from <https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines>
- Thurmair, G. (2009). Comparing different architectures of hybrid Machine Translation systems. Proceedings of MT Summit XII. Retrieved from <http://www.mt-archive.info/MTS-2009-Thurmair.pdf>
- Tripathi, S., & Sarkhel, J. K. (2010). Approaches to machine translation. *Annals of Library and Information Studies* (Vol. 57). Retrieved from <http://www.axistranslations.com/translation-article/machine-translation-definition.html>
- Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. *IFIP Congress*.
- Wagner, E. (1985). Post-editing systran - A challenge for commission translators. *Terminologie et Traduction*, 3. Advance online publication. doi:10.7202/003541ar
- Way, A. (2018). Quality expectations of machine translation. Retrieved from <https://arxiv.org/abs/1803.08409>
- Whale, S. (2017). AdaptiveMT, a self-learning machine translation engine: your questions answered - SDL Trados Blog. Retrieved August 20, 2018, from <https://blog.sdltrados.com/adaptivemt-self-learning-machine-translation-engine-questions-answered/>
- Wołk, K., & Marasek, K. Based on European Medicines Agency Leaflet Texts. (2015). Neural-based Machine Translation for Medical Text Domain. Based on European Medicines Agency Leaflet Texts. *Procedia Computer Science*, 64, 2–9. doi:10.1016/j.procs.2015.08.456
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., & Macherey, W. ... Dean, J. (2016). Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Retrieved from <https://ai.google/research/pubs/pub45610>

ADDITIONAL READING

Allen, J. (2003). Post-editing. In H. L. Somers (Ed.), *Computers and Translation: a translators guide* (pp. 297–317). John Benjamins., doi:10.1075/btl.35.19all.

Baker, M., & Saldanha, G. (2009). *Routledge encyclopedia of translation studies* (2nd ed.). Routledge., doi:10.4324/9780203872062.

Hutchins, J. (2007). *Compendium of translation software: commercial machine translation systems and computer-aided translation support tools. Mosaic A Journal For The Interdisciplinary Study Of Literature*. European Association for Machine Translation.

Hutchins, J. (2007b). Machine translation : A concise history. *Machine Translation*, 13(1 & 2), 1–21.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(8). Advance online publication. doi:10.1145/1380584.1380586

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2012). An Open-Source Toolkit for Integrating Shallow-Transfer Rules into Phrase-Based Statistical Machine Translation. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pp.41–54. Gothenburg, Sweden.

Shaw, F., & Gros, X. (2007). Survey of Machine Translation Evaluation.

Tertoolen, R. C. (2012). Desarrollos en la traducción automática: Esperando aún una traducción de alta calidad. *Journal of Chemical Information and Modeling*, 1(2). Advance online publication. doi:10.1017/CBO9781107415324.004

KEY TERMS AND DEFINITIONS

Adaptive Machine Translation: Machine translation systems whose architecture uses user feedback in order learn and to avoid repeating mistakes.

Corpus-Based Machine Translation: Machine translation systems whose architecture is based on the analysis of bilingual or multilingual corpora.

Hybrid Machine Translation: Machine translation systems whose architecture is based on the combination of the rule-based and the corpus-based approaches with the aim to combine the advantages of both approaches.

Machine Translation: The automatic process carried out by a software for translating text from one language to another.

Neural Machine Translation: Machine translation systems whose architecture is based on artificial neural networks.

Post-Editing: The revision process for a text that has been previously translated by an MT system.

Rule-Based Machine Translation: Machine translation systems whose architecture is based on linguistic information retrieved from dictionaries and grammars.

Translation Memory: A linguistic database that stores previously translated text so it can be reused for generating new translations.

Migrating Software Towards Mobile Technologies

7

Liliana Maria Favre

 <https://orcid.org/0000-0003-1370-1861>

Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina & Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina

INTRODUCTION

Today, the use of technology is central to the business world. Organizations need to redesign their processes and models so as not to be left out of the market in the near future. The adoption of new digital technologies offers a competitive advantage and performs as a market differentiator. In particular, new paradigms such as Pervasive Computing, Cloud Computing and the Internet of Things (IoT) are impacting the business world. Smartphones are at the core of these paradigms, letting us locate and easily interact with the world around us. With the Smartphone through Wi-Fi and the 5G, we will handle all kinds of objects connected to the network. Just as smartphones have already displaced the camera, the GPS, the music player, and the wallet, they will be on-ramp for a new IoT revolution. In this scenario, humans and things act synergistically as a whole and smartphones will continue to play a crucial role due to they are the main interface connecting people to the Internet (Stancovik, 2014) (Islam and Want, 2015). Smartphones are being transformed into a service center for different platforms in science, medicine, and education. Hossain et al. (2019) analyze the effects of variety-seeking intention by mobile phone usage on university students' academic performance.

In this context, most challenges for the competitive software industry are related to the problems caused by the proliferation of mobile platforms. New applications must support them as possible to remain profitable. Software applications can take full advantage of platforms only when they are built using native codebase. To address this problem a possible solution is to have different teams of developers who are fluent in a specific programming language to port an application to a specific platform. Instead of this traditional approach, organizations can use multiplatform or multi-paradigm cross-compiler based languages. The term “multiplatform” is used to refer source-source compilation, that is to say, the source code of these languages can be compiled into the source code of another programming language. Haxe is a good example of multiplatform languages. It allows using the same code to deploy an application on multiple platforms such as iOS, BlackBerry or Android. Specifically, it is an open-source high-level multiplatform programming language and compiler that can produce applications and source code for many different platforms from a single code-base (Haxe, 2019).

Frequently, the development of software component and applications aligned to mobile technologies requires adapting existing non-mobile software to different mobile platforms. For instance, there exist valuable software components and libraries implemented in C/C++ or Java that need to be adapted for mobile developments. There is the need to define systematic, reusable migration processes with a high degree of automation that reduce risks, time and costs of the cross-platform development. Novel technical frameworks for information integration, tool interoperability, and reuse can help to achieve these goals. Specifically, Model- Driven Engineering (MDE) is a software engineering discipline which emphasizes

DOI: 10.4018/978-1-7998-3479-3.ch061

the use of models and model transformations to raise the abstraction level and the degree of automation in software development (Brambilla et al., 2017). Productivity and some aspects of software quality such as maintainability or interoperability are goals of MDE.

Model Driven Developments (MDD) refer to forward engineering processes that use models as primary development artifacts. A specific realization of MDD is the Model-Driven Architecture (MDA) proposed by the Object Management Group (OMG) (MDA, 2019). Models, metamodels and model transformations play a major role in MDA. The MDA processes can be seen through a sequence of model transformations at different abstraction levels. A transformation is a process of converting a source model that conforms to a source metamodel in a target model that conforms to a target metamodel. The essence of MDA is Meta Object Facility (MOF), an OMG standard for defining metamodels that provides the ability to design and integrate semantically different artifacts in a unified way (MOF, 2016).

The OMG Architecture-Driven Modernization Task Force (ADMTF) is developing a set of specifications and promoting industry consensus on modernization. ADM is defined as “*the process of understand and evolve existing software assets for the purpose of software improvement, modifications, interoperability, refactoring, restructuring, reuse, porting, migration, translation, integration, service-oriented architecture deployment*” (ADM, 2019).

MDE progresses in the direction of achieving to develop software as reliable as any other engineering product, however, its impact on the software industrialization is limited. The main reasons are due, on the one hand, to a lack of repositories that contain reliable metamodels of the programming languages, and transformations between the different metamodels. It is worth considering that, MDE developers, instead of needing CASE tools to support the development on specific platforms, need to reuse metamodels and transformations. The incipient degree of progress in the specification of them requires that the developer, in general, must build them from scratch. This implies additional costs and time involved in these activities that must be compared with the total benefit. In addition, it is important to note that the construction of a metamodel should be considered critical development since the instances of a metamodel (models) must be reliable but currently, it is not focused in this way.

On the other hand, Model-driven approaches were broadcast emphasizing on processes viewed as a sequence of transformations between metamodels defined through the official standards. However, it is convenient to have a more flexible vision of MDE that allows integrating software developed from other frameworks or tools that have been tested through its use for several years and software developers prefer to continue using them. For example, software developers prefer to use multiplatform languages over model-driven developments that require defining metamodels and transformations for all multiple mobile platforms. Integrating multiplatform languages to MDE developments for the modernization of software would not imply defining metamodels and transformations for all the platforms it supports, but only for the multiplatform language.

Metamodeling is crucial in Model-Driven processes. It is important to formalize and reason about MOF metamodels and we propose to exploit the strong background achieved by the community of formal methods. Formal methods provide systematic and rigorous techniques to reduce ambiguities and inconsistencies in software development. Within the next few decades, tools based on verification will be as useful and widespread for software development as they are today in critical systems (Beckert and Hahnle, 2014). Behind the specifications and formal methods exists a community that has developed theories, languages, methods, and tools, but the software industry does not take advantage of them, in particular, to effectively address the software crisis. There is already a substantial body of research, tools and case studies demonstrating that it is possible to develop software as reliable as any other engineering product. To enhance the software industry, artifact analysis tools that support novel combinations

of code analysis techniques, model checking, testing and theorems provers are emerging (Beckert & Hahnle, 2014).

In this chapter, we describe a lightweight MDE approach based on the ADM standards. In this context, a lightweight process is the one that leaves open the possibility to fall back to the traditional methodology without model-driven standards. It is based on the integration of cross-platform frameworks, and formal methods with ADM. First, an infrastructure for the definition of reliable metamodels that integrates semi-formal and formal specifications based on algebraic specifications of metamodels is analyzed. Next, an integration of ADM standards with cross-platform development for the migration of desktop software to mobile platforms through the multiplatform language Haxe is described. A repertory of new metamodels and transformation necessary to adapt object-oriented software to different mobile platform through Haxe is presented. Besides, the chapter compares MDE software migration versus traditional migration processes as well as it describes the existing challenges to achieve the results in real projects in the scientific and industrial field.

BACKGROUND

Model Driven Engineering (MDE) is a software development approach that focuses on the use of models to define software development processes with a high degree of automation starting from reusable transformations between models that conform to metamodels. The underlying idea in MDE is that software development can be broken down into standardized, highly automated activities that can mass produce software applications reducing costs, development time and risks.

Different acronyms are associated with model-driven developments: MBE (Model Based Engineering), MDE (Model Driven Engineering), MDD (Model Driven Development), MDA (Model-Driven Architecture), MDSM (Model Driven Software Modernization) and ADM (Architecture Driven Modernization).

MBE is the branch of software engineering in which software models play an important role, being the basis of development. However, there is no direct link between the models and the generated software precisely defined through transformations.

MDE can be viewed as a subset of MBE. It is the branch of software engineering in which processes are driven by models, i.e. models are the primary artifacts of different software processes. MDE has emerged as a new software engineering discipline which emphasizes the use of models and model transformations to raise the abstraction level and the degree of automation in software development. Productivity and some aspects of the software quality such as maintainability or interoperability are goals of MDE.

MDD refers to forward engineering processes that use models as primary development artifacts. A specific realization of MDD is MDA (MDA, 2014). The outstanding ideas behind MDA are, separating the specification of the system functionality from its implementation on specific platforms, managing the software evolution from abstract models to implementations. Models play a major role in MDA, which distinguishes at least:

- Computation Independent Model (CIM), a model that describes a system from the computation independent viewpoint
- Platform-Independent- Model (PIM), a model with a high level of abstraction that is independent of any independent technology,
- Platform-Specific-Model (PSM) a tailored model to specify the system in terms of the implementation constructs available in one specific platform

The Unified Modeling Language (UML) combined with the Object Constraint Language (OCL) is the most widely used way for writing PIM and PSMs.

An MDA process focuses on the automatic transformation of different models that conform to MOF metamodel, the standard for defining metamodels in the context of MDA (MOF, 2016). It provides the ability to design and integrate semantically different languages such as general-purpose languages, DSLs and modeling languages in a unified way. MOF can be considered the essence of MDA allowing different kinds of artifacts from multiple technologies to be used together in an interoperable way. MOF provides two metamodels EMOF (Essential MOF) and CMOF (Complete MOF). EMOF favors the simplicity of implementation over expressiveness. The OMG standard related to model transformation is the MOF 2.0 Query, View, Transformation (QVT) metamodel (QVT, 2016).

In the context of ADM, a set of modernization specifications is developed. ADMTF is developing a set of standards of which we are interested in KDM (Knowledge Discovery Metamodel) and ASTM (Abstract Syntax Tree Metamodel) (ASTM, 2019) (KDM, 2019). KDM is a metamodel for knowledge discovery in software that allows representing information related to existing software assets, their associations, and operational environments regardless of the implementation programming language and runtime platform. KDM is the foundation for software modernization representing entire enterprise software systems, not just code. ASTM is a specification for modeling elements to express abstract syntax trees. KDM and ASTM are two complementary modeling specifications. KDM establishes a specification that allows representing semantic information about a software system, whereas ASTM establishes a specification for representing the source code syntax by means of AST (Abstract Syntax Tree) in a representation that is sharable among multiple tools. ASTM acts as the lowest level foundation for modeling software within the OMG ecosystem of standards, whereas KDM serves as a gateway to the higher-level OMG models.

MDSM is a particular form of reengineering for the technological and functional evolution of legacy systems that begins to be identified in the early 21st century (Brambilla et al., 2017). It is based on model-driven processes of reverse engineering, restructuring and forward engineering. Software reengineering distinguishes different levels of analysis and provides foundations for logical transformations at different abstraction levels, especially for transformations that can be of different types: T2M (Text-to-Model), M2M (Model-to-Model), or M2T (Model-to-Text).

The Eclipse Modeling Framework (EMF) was created for facilitating system modeling and the automatic generation of Java code (EMF, 2019) (Steinberg et al., 2009). EMF started as an implementation of MOF resulting Ecore, the EMF metamodel comparable to EMOF. EMF has evolved starting from the experience of the Eclipse community to implement a variety of tools and to date is highly related to MDD. In this context, the subproject Model to Model Transformation (MMT), hosts model-to-model transformation languages. Transformations are executed by transformation engines that are plugged into the Eclipse Modeling infrastructure. For instance, Atlas Transformation Language (ATL) is a model transformation language and toolkit that provides ways to produce a set of target models from a set of source models (ATL, 2019). Another subproject is Acceleo, which is an implementation of the Model-to-Text (M2T) transformation standard of the OMG for EMF-based models (Acceleo, 2019). Acceleo is used in forward engineering processes.

Today, the most complete technology that support ADM is MoDisco, which provides a generic and extensible framework to facilitate the development of tools to extract models from legacy systems and use them on use cases of modernization. As an Eclipse component, MoDisco can integrate with plugins or technologies available in the Eclipse environment (Modisco, 2019) (Bruneliere et al., 2014). In particular, the ASTM specification mainly includes the following metamodels:

- Generic Abstract Syntax Tree Metamodel (GASTM), a generic set of language modeling elements common across numerous languages establishes a common core for language modeling.
- Language Specific Abstract Syntax Tree Metamodels (SASTM) for particular languages such as Ada, C, Fortran and Java, that are modeled in MOF or MOF compatible forms.
- Proprietary Abstract Syntax Tree Metamodels (PASTM), that express ASTs for languages such as Ada, C and COBOL modeled in formats that are not consistent with MOF, the GSATM, or SASTM and define the minimum conformance standards needed to support model interchange.

In particular, C++ is one of the most commonly used programming language in science and engineering domains and numerous legacy software components written in C++ require to be modernized. EMF4CPP is the first step at providing a set of tools for MDD in C++ as an alternative to the Eclipse tools for Java (EMF4CPP, 2016). It is a C++ implementation and type mapping for the EMF core, the Ecore metamodel. The main facilities provided by EMF4CPP are to generate C++ code from Ecore metamodels and to parse and serialize models and metamodels from and into XMI documents (Jäger et al., 2016). However, an implementation of a MOF-compliant C++ metamodel is necessary for other MDE processes, in particular ADM (e.g., reverse engineering or software modernization).

Multiplatform Development

Today, one of the major challenges for software developers is dealing with the rapid proliferation of mobile platforms that entails the high cost, technical complexity, and risk of targeting development to a wide spectrum of platforms. Taking into account the diversity of platforms, which are also in a dizzying evolution, it is practically intractable defining as many metamodels and transformations as mobile platforms exist. To address this problem a possible solution is to have different teams of developers who are fluent in specific programming language to port an application to a specific platform. Instead of this traditional approach, organizations can use multiplatform or multi-paradigm cross-compiler based languages. In this context, the term “*multiplatform*” is used to refer source-source compilation, that is to say, the source code of these languages can be compiled into source code of other programming language. In this direction, the Haxe language emerges as an open-source high-level multiplatform programming language and compiler that can produce applications and source code for many different platforms from a single code-base.

The Haxe programming language is a high level programming language that mixes features of object oriented languages and functional ones. It is similar (but not pure) to object-oriented languages. Haxe includes a set of common functions that are supported across all platforms, such as numeric data types, text, arrays, binary and some common file formats (Haxe, 2019).

To date, Haxe supports nine target languages which allow for different use-cases: JavaScript, Neko, PHP, Python, C++, ActionScript3, Flash, Java and, C# and test your app on each platform. The Haxe language can be used in a variety of domains such game development, web development and mobile development. In particular, it facilitates mobile development sharing code between key platforms and accessing native functionality without sacrificing performance. The idea behind Haxe is to allow developers choose the best platform for a specific development. To achieve this, it provides a standardized language, a standard library that works the same on all platforms and platform specific libraries that allow accessing the full API for a given platform from Haxe.

Related Work

The first results that address MDE and mobile developments are specifically linked to MDD. A survey of MDD approaches for mobile applications is presented in (Umuhoza and Brambilla, 2016). Various authors describe the challenges of mobile software development. A DSL (Domain Specific Language), named MobDSL, to generate applications for multiple mobile platforms is described in (Kramer, Clark and Oussena, 2010). They perform the domain analysis on two cases in the Android and iPhone platforms. This analysis allows inferring the basic requirements of the language defined by MobDSL. From another angle, Dehlinger and Dixon (2011) highlight creating user interfaces for different kinds of mobile devices, providing reusable applications across multiple mobile platforms, designing context-aware applications and handling their complexity and, specifying requirements uncertainty. On the other hand, a proposal for supporting mobile application development by using models as inputs to an emulator is outlined at (Bowen and Hinze, 2011). The authors describe an MDD-based emulator for using in the design of graphical interfaces and interactions. They propose to transform functional behavior and requirement models with design restrictions into emulated applications. From another perspective, Acerbis et al. (2015) describe a comprehensive tool suite called WebRatio Mobile Platform for model-driven development of mobile applications. It is based on an extended version of OMG standard language called IFML (Interaction Flow Modeling Language) empowered with primitives tailored to mobile systems that enable specification of mobile-specific behaviors.

The following authors show integration of reverse engineering MDE. Favre (2010) describes a reverse engineering approach that fits with MDD. A framework to integrate different techniques that come from compiler theory, metamodeling and formal specification is presented. It emphasizes the use of static and dynamic analysis for generating MDA models. Fernández Candel et al. (2019) describe a model-driven reengineering approach for migrating PL/SQL triggers to Java-based on KDM to represent legacy code and traditional techniques to complement ADM. Test Driven Development (TDD) is used to incrementally develop model transformations that were written in Java instead of using some transformation language like ATL or QVT imperative. A systematic literature review on model-driven reverse engineering approach may be found at (Raibulet et al., 2017)

The following authors relate smartphones and mobile development with new technologies such as Cloud Computing and IoT. Islam and Want (2014) describes major trends affecting future smartphone design and use such as personal computers, the IoT, multimedia delivery, low power operation, wearable computing and context awareness. On the other hand, Ejarque, Miccsik and Badia (2015) present a solution for facilitating the migration of applications to the cloud, inferring the most suitable deployment model for the application and automatically deploying it in the available Cloud providers. The findings of experiments carried out to understand the impact of application characteristics, cloud and architecture and the android emulator used, on application performance when the application is augmented to the cloud, are shown at (Joshi et al., 2015). Gonzalez Garcia et al. (2015) define a domain-specific language (DSL) that allows specifying the coordination and communication between different types of smart objects.

Several works highlight research directions and challenges on the IoT. The spectrum of research needed to achieve IoT on a large scale requires research along with many different directions. Stankovic (2014) identifies five prominent research communities that involve the smart vision of the world: IoT, mobile computing, pervasive computing, wireless sensors, networks, and cyber-physical systems. Zanella et al. (2014) present a discussion of the IoT for smart cities. The authors describe a general reference framework for the design of an urban IoT. An implementation of an urban IoT that has been realized in the city of Padova is described. Key research topics are enumerated and research problems within these

topics are discussed. Miranda et al. (2015) propose an infrastructure that supports the evolution from the “Internet of Things” to the “Internet of People” (IoP). The IoP is used in the sense of bringing the IoT closer to people, for easily integrate them and exploit its benefits.

The following articles are directly related to the chapter proposal. A reengineering process that integrates traditional reverse engineering techniques such as static and dynamic analysis with MDA is presented at (Améndola and Favre, 2013). The article describes a case study that shows how to move CRM (Customer Relationship Management) applications from desktop to mobile platforms. The proposal was validated in the open source application platform Eclipse, EMF, EMP, ATL, and Android platform. Diaz Bilotto and Favre (2016) describe a migration process from Java to mobile platforms through the multiplatform language Haxe. A framework for modernizing desktop software based on MDE is presented at (Favre, 2018).

INTEGRATING MDE WITH FORMAL METHODS AND CROSS-PLATFORM DEVELOPMENT

This section describes the main focus of the article: the construction of metamodels starting from the integration of MOF-compliant metamodels and formal specification, and a migration process of non-mobile software to different mobile platforms based on MDE, ADM, and Cross-Platform Development.

Formalizing MOF-Compliant Metamodels

An essential aspect to apply this approach is to have reliable definitions of metamodels. It is important to formalize and reason about MOF metamodels and we propose to exploit the strong background achieved by the community of formal methods. It is important to remark that the instantiation of a metamodel produces models, which in turn are instantiated. So, having errors in a metamodel leads to having errors in its model instances. Besides, a model can be well-formed but still be incorrect. A combination of MOF metamodeling and formal specification can help metadesigners to address these issues.

A major obstacle for specifying metamodels MOF or Ecore is that they must master two different languages: UML for expressing class diagram that capture the structure of the domain and the OCL language for expressing associated semantic (OCL, 2014). Cadavid et al. (2015) observe that, in general, the metamodels concentrate the specifications in a subset of concepts, most of them are loosely coupled to the underlying structure. Most MDE semiformal metamodels do not have support for typing metamodels and the notion of polymorphism at the metamodel level is imprecise due to the type of a metamodel is not defined. These are important limitations regards to the generation of valid instances to test models transformations and to reuse models, metamodels and model transformations.

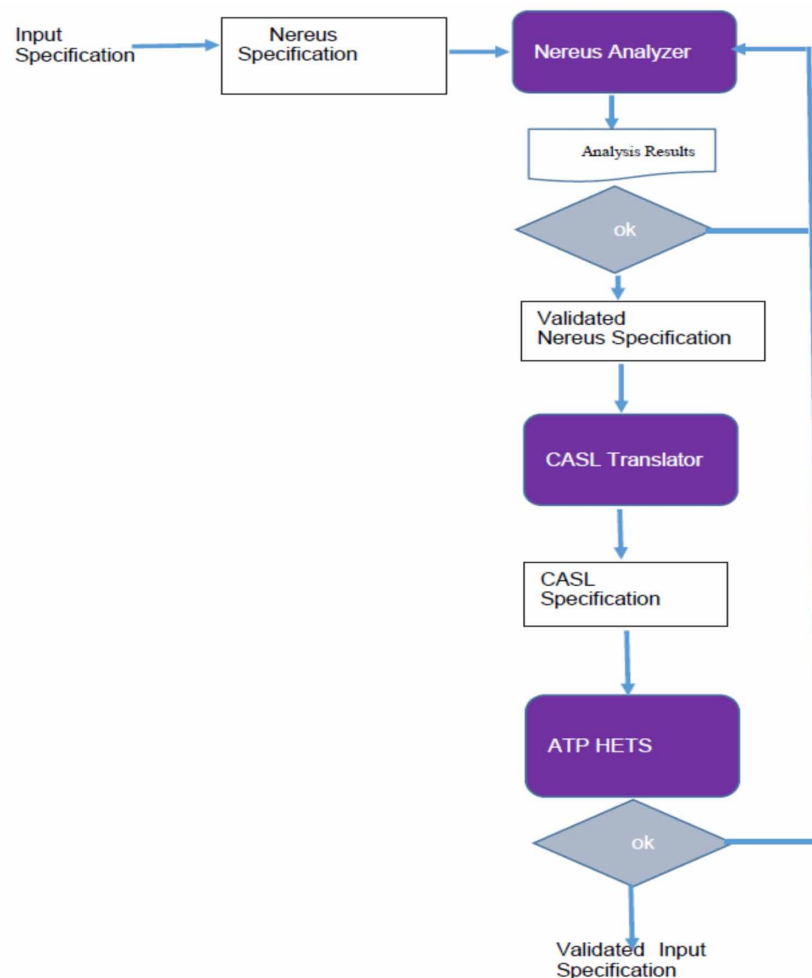
To address these problems, a formal metamodeling language called Nereus was defined. Formal methods offer rigor and precision while reducing ambiguity and inconsistency. Nereus allows specifying MOF or Ecore metamodels and reasoning about them. A detailed description of Nereus may be found at (Favre, 2009) (Favre and Duarte, 2016). A set of Nereus tools to make formal metamodeling feasible in practice was developed. The semantic of Nereus was given in the CASL language that can be linked with Automatic Theorem Provers (ATP) provided by Hets (Mossakowski, Maeder, and Codescu, 2019).

The process of meta-model construction is based on the analysis of fragments of models (examples of concrete instances) from which a metamodel is induced. These fragments focus on some interest aspects of the metamodel. In a first step, they are represented as an Ecore metamodel complemented with

textual specifications in *OCLinEcore* Editor that overcomes some limitations of the Ecore Editor, for instance, Ecore does not detect syntactic and semantic errors in OCL constraints. Another advantage is that *OCLinEcore* works with text representations instead of using *XMI* (XML Metadata Interchange), allowing greater modifiability, readability and better integration with versioning tools. A metaclass is created for each object of a different type (if it does not exist) and the attributes and operations and their restrictions are identified. Then, the different relationships between metaclasses are identified. The metamodel is subsequently transformed into a Nereus metamodel. The formal specification is analyzed by using the analyzer of Nereus and is modified according to the results of the translation process with the goal of obtaining a syntactically correct specification that, through its integration with CASL could be formally validated.

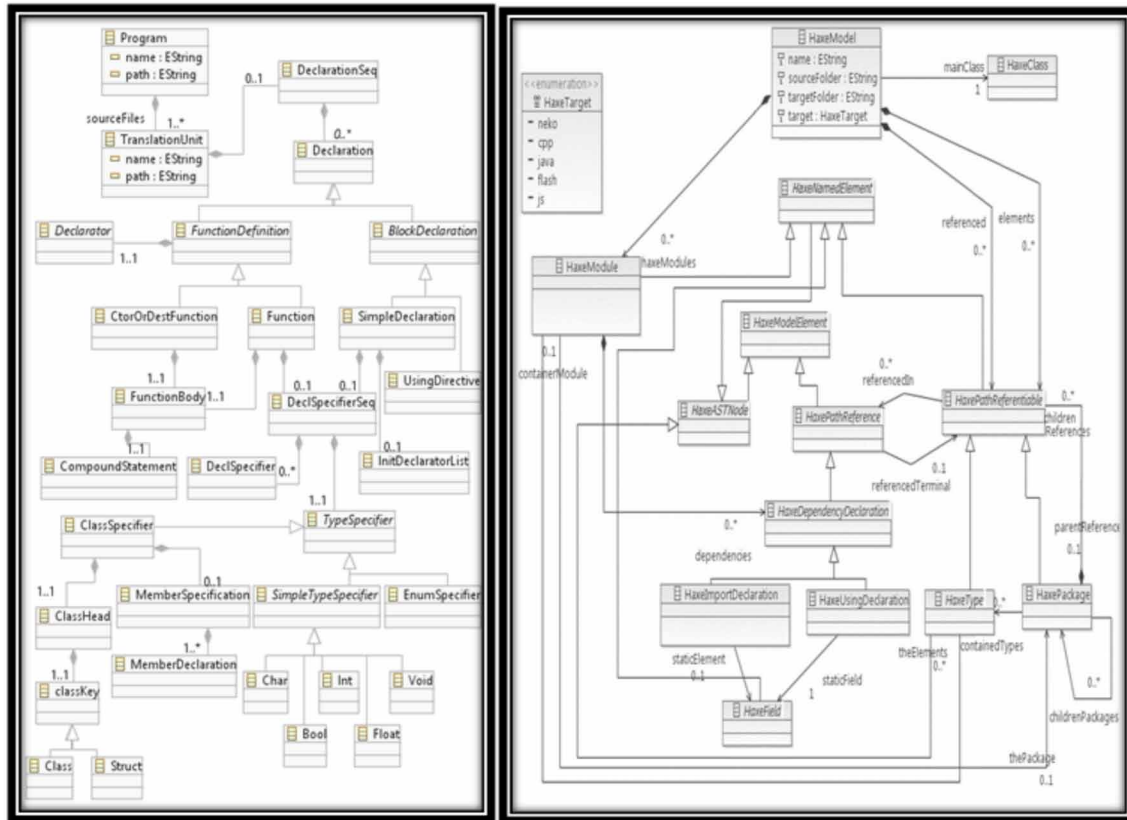
The typical flow with formal tools that is applied to analyze the specifications of metamodel is shown in Figure 1. In the first step, a semiformal metamodel is transformed into a Nereus specification. The formal specification is analyzed and modified according to the results of the translation process. The goal is to obtain a syntactically correct specification that, through its integration with CASL language can be formally validated. Nereus allows us to experiment both defining metamodels for DSLs from scratch, and analyzing fragments of already defined metamodels, detecting inconsistencies and proposing solutions.

Figure 1. Constructing metamodels



In the context of this work, the Haxe metamodel and C/C++ metamodel were defined. Figure 2 partially shows the C / C++ metamodel (on the left) and the Haxe metamodel (on the right).

Figure 2. C/C++ and Haxe Metamodels



The C++ metamodel conforms to Ecore. The root metaclass is *Program* that represents a C++ program, which owns source files, instances of *TranslationUnit*. A translation unit contains declarations such as block declaration, function definitions, template declarations, among others. A *SimpleDeclaration*, instance of *Block-Declaration*, has a *DeclSpecifierSeq* that is a sequence of *DeclSpecifiers* which refers to a declaration specifiers and a type specifier. In addition, a simple declaration has an *InitDeclaratorList* containing a variable declaration list that is a list of specifiers and the name of a variable and its corresponding initialization. A *FunctionDefinition* has a *Declarator* containing the function identifier and the parameter list. *Function* and *CtorOrDestFunction*, instances of *FunctionDefinition*, have a body that contains compound statements such as declarations, iterations, and selections. In addition, a *Function* has a *DeclSpecifierSeq* that is a sequence of *DeclSpecifiers* such as function specifiers and a type specifier. *TypeSpecifier* subclasses are *SimpleTypeSpecifier*, *ClassSpecifier* and *EnumSpecifier* among others. A *ClassSpecifier* has a *ClassHead* containing the class key (class or struct) and a *MemberSpecification* that contains *MemberDeclarations* such as variables, function declarations, function definitions, constructors, destructor and template members.

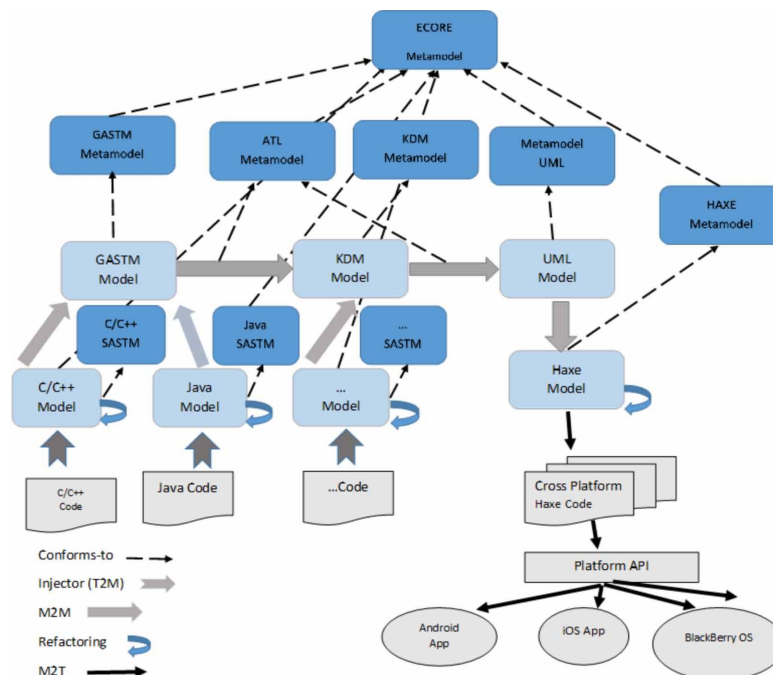
Also, the Haxe metamodel was defined in Ecore. The main metaclasses of the Haxe metamodel are those that allow specifying an application using Haxe as language. One of the main metaclasses of the

metamodel is *HaxeModel* that serves as element container used to describe an application and store additional information on it, for example, some options of compilation and different metaclasses for modeling such as modules, classes and packages. The metaclasses used directly by *HaxeModel* are *HaxeModule* and *HaxePathReferentiable*. Starting from the relations *HaxeModules*, *referenced* and *elements*, the class *HaxeModel* allows storing different information. Relation *HaxeModules* allows accessing the different Haxe modules used in the project. Through relation *elements*, it is possible to access the different elements of the package tree. Relation *referenced* provides access to elements which are referenced in the project but are not defined completely. In the case of relations and referenced elements, the type used is *HaxePathReferentiable*, which is the parent type of metaclasses such as *HaxeType* and *HaxePackage*. The Haxe language includes different kind of types such as class (the type's class and interface), function, abstract type, enumeration, and anonymous structures. A full description of both, the C/C++ and Haxe metamodels may be found at (Duthey and Spina, 2016).

Migrating Software to Mobile Platforms

A process to migrate legacy code to different mobile platforms based on ADM standards and cross-platform development is described. The migration process follows model-driven development principles: all artifacts involved in the process can be viewed as models that conform to Ecore meta-metamodel, the process itself can be viewed as a sequence of model-to-model transformations and the extracted information is represented in a standard way through Ecore metamodels. For each transformation, source and target metamodels are specified. A source metamodel defines the family of source models to which transformations can be applied. A target metamodel characterizes the generated models. Different kind of transformations such as T2M, M2M, and M2T are distinguished in the steps of reverse engineering, refactoring/restructuring, and forward engineering. Figure 3 summarizes the proposed process.

Figure 3. The migration process



The first step is the reverse engineering of source code to obtain the abstract syntax tree of the code and consists of two stages:

- Generating the first model of the code by using a model injector. This model conforms to the source code metamodel, such as C++ and Java. The obtained model could be refactored to reorganize and modify the syntactic elements and to improve the design. The refactoring is implemented as a model-to-model transformation whose source and target models are instances of source code metamodel.
- Generating the abstract syntax tree model, instance of the GASTM metamodel, from the model obtained in the previous stage by an ATL model-to-model transformation.

In this first step of the process, an injector and a transformation to obtain the GASTM model must be implemented for each language, whereas the sequence of transformations involved in the followings steps of the migration process is independent of the language of the legacy code.

The second step generates the KDM model. This process is carried out by means of an ATL model-to-model transformation that takes as input a model conforming to the GASTM metamodel and produces a model conforming to the KDM metamodel.

The advantage of this intermediate step is that starting from the KDM model it is possible to obtain high-level models such as UML class diagrams, activity diagrams and use cases diagrams. These models could be refactored and be the starting point for generating code. This step is common for each source language.

UML models generated from KDM are transformed to UML models for the Haxe platform from which it is possible to generate Haxe code using M2T transformations expressed in Acceleo (Acceleo, 2019).

Given that Modisco provides T2M transformations for Java, we will discuss the stages for the migration of C++ code to different mobile platforms.

The first transformation extracts an AST model specific to C++ from code. To carry out this task, we constructed a model injector by using EMFText (EMFText, 2019). To generate this injector, EMFText requires the language metamodel and the concrete syntax specification. In our approach, to generate the injector we first specified the C++ metamodel based on the C++ grammar. Then, we specified the concrete syntax that defines the textual representation of all metamodel concepts. Taking these specifications, the EMFText generator derives an advanced text editor that uses a parser and printer to parse language expressions to EMF models or to print EMF models to languages expressions respectively.

The second transformation takes as input the model obtained in the previous step and release a generic AST model conforming to the GASTM metamodel. This transformation specifies the way to produce GASTM projects (target) from C++ programs (source). The previous transformations are dependent on the legacy code language, that is, the model injector and the transformation to obtain the generic AST model are dependent on C++. In contrast to the previous stage, the sequence of transformations from GASTM models to Haxe models is independent of the language.

Haxe allows writing mobile applications that target all major mobile platforms in a straightforward way. The generated code is syntactically correct, although, it does not compile on other platforms without doing changes due to the code refers to proprietary technologies of C++. To run on mobile environments, these technologies can be replaced with OpenFL and HaxeUI (that is an open-source, multi-platform application-centric user interface framework designed for Haxe and OpenFL) (OPEN FL, 2019)

SOLUTIONS AND RECOMMENDATIONS

The previous sections describe two main contributions:

- The migration of object-oriented software to various platforms based on the integration of ADM with Haxe.
- The construction of metamodels based on the integration of semiformal and formal specifications.

With respect to the first contribution, it is possible to migrate the logic of software applications and run them on mobile platforms without having to completely rewrite them. Due to the fact that the objective of the migration is not only “compile” an application in a mobile platform but also to create a fully functional version of the application using quality criteria, small adjustments had to be made.

MDE approach based on ADM provides benefits with respect to ad-hoc migration. One of the benefits is to increase productivity in software development due to the automation that is introduced in the generation of artifacts. In MDE migration projects is more cost-effective when the same process must be repeated frequently as the case of migrations to different platforms.

Another benefit of the process is its independence from source and target technologies. In our approach, the intermediate models act as decoupling elements between source and target technologies. Independence is achieved with injectors and, M2M and M2T transformations. Besides, in a transformation sequence, models could be an extension point to incorporate new stages, for instance, to connect with different multiplatform languages. Model transformations allow developers to concentrate on conceptual aspects of the relations between models and then to delegate most of the migration process to the transformation rules, whereas in the brute-force redevelopment, developers need to migrate by hand the legacy systems, making over and over again the same task.

In the context of our proposal, Haxe allows using the same code to deploy an application on multiple mobile platforms. The migration of object-oriented software to mobile platforms is achieved through the definition of a metamodel for Haxe and a single M2M transformation, delegating the integration with different mobile platforms to Haxe, that is, the evolution of existing platforms or the integration of new mobile platforms is considered a goal of Haxe.

The following limitations our approach were observed. The set of metamodels and discovers that can be used is limited. In our approach, it was necessary to define metamodels for the C / C ++ and Haxe languages and, a discoverer for C/C ++. However, the cost of preliminary activities could be recovered in later developments.

With respect to the second contribution, a metamodeling framework based on MOF and the algebraic formalism that focuses on automatic proofs and tests was provided. The central components of our approach are the definition of the formal language Nereus closed to MOF and the development of tools for formal metamodeling: the Nereus analyzer and the Nereus-to-CASL translator.

FUTURE RESEARCH DIRECTIONS

Figure 3 shows the essential activities to migrate non-mobile software to various mobile platforms. The different stages were validated in medium-scale developments. However, each stage requires more research and experimentation to show the real impact on larger scale developments.

A semiautomatic tool that takes as its starting point a C++ project and produces one written in Haxe, in addition to the generated intermediate files, was developed to prove model transformations. The tool is semi-automatic because it requires additional actions on the Haxe code to make it fully functional (Duthey and Spina, 2016). It is also still a challenge to migrate the visual part if the graphical interfaces are linked to specific libraries. The tool has been used in the migration of games such as Tetris and medium-scale software applications.

Also, it remains to analyze the generation of Haxe code from UML models more deeply. Considering that from KDM it is possible to generate UML models at the PIM level and that currently, CASE MDA / UML tools do not support multiplatform languages it is necessary to move from the stage of validation of transformations through prototypes to a real integration with existing CASE tools, either commercial or free.

An important contribution was to achieve an infrastructure to define metamodels rigorously. It remains to be completed the integration between formal metamodels and Ecore metamodels through the automatic generation of Ecore metamodels from Nereus metamodels.

CONCLUSION

An ADM approach for migrating non-mobile software to different mobile platforms was described. It takes advantage of tools emerged in the cross-platform and formal methods communities. The transformation to different platforms is achieved with the definition of a metamodel for Haxe. It is important to note that was not necessary to define metamodels for all the languages and platforms supported by it, as it would have been necessary in order to realize a “pure” MDE development.

An infrastructure for the definition of metamodels based on the integration of semi-formal and formal specifications was described. It was validated with the definition of two metamodels: the Haxe metamodel and C/C++ metamodel, previously non-existent.

Another contribution to the MDE community is the definition of a set of T2M, M2M and M2T transformations that support the migration process and can be reused in other MDE processes. Considering that mobile application developers need frequently adapt software components and applications developed in languages such as Java or C/C++, the model-driven migration processes could be reused and the cost of preliminary activities would be recovered.

REFERENCES

- Acceleo. (2019). *Obeo. Acceleo Generator*. Retrieved July 25, 2019 from <http://www.eclipse.org/Acceleo/>
- Acerbis, R., Bongio, A., Brambilla, M., & Butti, S. (2015). *Model-Driven Development Based on OMG's IFML with WebRatio Web and Mobile Platform*. In *Engineering the Web in the Big Data Era* (Vol. 9114). Springer-Verlag.
- ADM. (2019). *Architecture-driven modernization task force*. Retrieved July 25, 2019 from <http://www.adm.org>
- Améndola, F., & Favre, L. (2013). Adapting CRM Systems for Mobile Platforms: An MDA Perspective. *International Journal of Computer & Information Science*, 14(1), 31–40.

ASTM. (2011). *Abstract Syntax Tree Metamodel*, version 1.0, OMG Document Number: formal/2011-01-05. Retrieved July, 25, 2019 from <http://www.omg.org/spec/ASTM>

ATL. (2019). *Atlas Transformation Language Documentation*. Retrieved July 25, 2019 from <http://www.eclipse.org/atl/documentation/>

Beckert, B., & Hahnle, R. (2014, Jan.). Reasoning and Verification: State of the Art and Current Trends. *IEEE Intelligent Systems*, 20–29.

Bowen, J., & Hinze, A. (2011). Supporting mobile application development with model-driven emulation. *Journal of the ECEASST*, 45, 1–5.

Brambilla, M., Cabot, J., & Wimmer, M. (2017). *Model-Driven Software Engineering in Practice*. In *Synthesis Lectures on Software Engineering* (2nd ed.). Morgan & Claypool Publishers.

Bruneliere, N., Cabot, J., Dupé, G., & Madiot, F. (2014). MoDisco: A Model Driven Reverse Engineering Framework. *Information and Software Technology*, 56(8), 1012–1032. doi:10.1016/j.infsof.2014.04.007

Cadavid, J., Combemale, B., & Baudry, B. (2015). An analysis of metamodeling practices for MOF and OCL. *Computer Languages, Systems & Structures*, 41, 42–65. doi:10.1016/j.cl.2015.02.002

Dehlinger, J., & Dixon, J. (2011). Mobile application software engineering: Challenges and research directions. In *Proceedings of the Workshop on Mobile Software Engineering* (pp. 29–32). Berlin: Springer-Verlag.

Diaz Bilotto, P., & Favre, L. (2016). Migrating JAVA to Mobile Platforms through HAXE: An MDD Approach. In A. M. Cruz & S. Paiva (Eds.), *Modern Software Engineering Methodologies for Mobile and Cloud Environments* (pp. 240–268). IGI Global. doi:10.4018/978-1-4666-9916-8.ch013

Duthey, M., & Spina, C. (2016). *Migrating C/C++ to mobile platforms through MDD* (Undergraduate Thesis). Computer Science Department, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina.

Ejarque, J., Micsik, A., & Badia, R. (2015). Towards Automatic Application Migration to Clouds. In *Proceedings IEEE 8th Int. Conf. on Cloud Computing (CLOUD)* (pp. 25–32). New York: IEEE.

EMF4CPP. (2016). *What is EMF4CPP?* Retrieved July 25 from <https://code.google.com/archive/p/emf4cpp/>

EMF. (2019). *Eclipse Modeling Framework (EMF)*. Retrieved July 25, 2019 from <http://www.eclipse.org/modeling/emf/>

EMFText. (2019). Retrieved July 25, 2019 from www.emftext.org

Favre, L. (2009). A Formal Foundation for Metamodeling. In *Proceedings of 14th Ada-Europe International Conference. Lecture Notes in Computer Science* (Vol 5570, pp.177–191). Berlin: Springer-Verlag. 10.1007/978-3-642-01924-1_13

Favre, L. (2010). *Model Driven Architecture for Reverse Engineering Technologies: Strategic Directions and System Evolution*. Hershey, PA: IGI Global. doi:10.4018/978-1-61520-649-0

Favre, L. (2018). A Framework for Modernizing Non-Mobile Software: A Model-Driven Engineering Approach. In *Protocols and Applications for the Industrial Internet of Things* (pp. 192–224). IGI Global. doi:10.4018/978-1-5225-3805-9.ch007

Favre, L., & Duarte, D. (2016). Formal MOF Metamodeling and Tool Support. In *Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development. MODELSWARD 2016* (pp. 99–110). Roma, Italy: SCITEPRESS.

Fernández Candel, C. J., García Molina, J., Bermúdez Ruiz, F. J., Hoyos Barceló, J. R., Sevilla Ruiz, D., & Cuesta Viera, B. J. (2019). Developing a model-driven reengineering approach for migrating PL/SQL triggers to Java: A practical experience. *Journal of Systems and Software*, 151(May), 38–64. doi:10.1016/j.jss.2019.01.068

González García, C., & Espada, J. (2015). MUSPEL: Generation of Applications to Interconnect Heterogeneous Objects Using Model-Driven Engineering. In V. G. Díaz, J. M. C. Lovelle, & B. C. P. García-Bustelo (Eds.), *Handbook of Research on Innovations in Systems and Software Engineering* (pp. 365–385). IGI Global. doi:10.4018/978-1-4666-6359-6.ch015

Haxe. (2019). *The Haxe Language*. Retrieved July 25, 2019 from <https://Haxe.org/>

Hossain, S. F. A., Nurunnabi, M., Hussain, K., Saha, S. K., & Wang, S. (2019). Effects of variety-seeking intention by mobile phone usage on university students' academic performance. *Cogent Education*, 6(1), 1574692. doi:10.1080/2331186X.2019.1574692

Islam, N., & Want, R. (2014). Smartphones: Past, present and future. *Pervasive Computing*, 13(4), 82–92. doi:10.1109/MPRV.2014.74

Jäger, S., Maschotta, R., Jungebloud, T., Wichmann, A., & Zimmermann, A. (2016). An EMF-like UML Generator for C++. In *Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development (MODELSWARD 2016)*, (pp. 309–316). SCITEPRESS. 10.5220/0005744803090316

Joshi, P., Nivangune, A., Kumar, R., Kumar, S., Ramesh, R., Pani, S., & Chesum, A. (2015). Understanding the Challenges in Mobile Computation Offloading to Cloud through Experimentation. In *2nd ACM Int. Conf. on Mobile Software Engineering and Systems (MOBILESoft)*, (pp. 158–159). Florence, Italy: ACM. 10.1109/MobileSoft.2015.43

KDM. (2016). *Knowledge Discovery Meta-Model (KDM)*, *OMG Document Number: formal/2016-09-01*. Retrieved July, 25, 2019 from <http://www.omg.org/spec/KDM/1.4>

Kramer, D., Clark, T., & Oussena, S. (2010). MobDSL: A domain specific language for multiple mobile platform deployment. In *Proceedings of IEEE Int. Conf. on Networked Embedded Systems for Enterprise Applications (NESEA)*, (pp. 1–7). Perth, Australia: IEEE. 10.1109/NESEA.2010.5678062

MDA. (2019). *The Model-Driven Architecture*. Retrieved July 25, 2019 from www.omg.org/mda/

Miranda, J., Makitalo, N., Garcia-Alonso, J., Berrocal, J., Mikkonen, T., Canal, C., & Murillo, J. (2015). From the Internet of Things to the Internet of People. *IEEE Internet Computing*, 19(2), 40–47. doi:10.1109/MIC.2015.24

MoDisco. (2019). *Model Discovery*. Retrieved July 25, 2019 from <https://eclipse.org/MoDisco/>

MOF. (2016). *Meta Object Facility (MOF) Core Specification, Version 2.5, OMG Document Number: formal/2016-11-01*. Retrieved July 25, 2019 from <http://www.omg.org/spec/MOF/2.5.1/>

Mossakowski, T., Maeder, C., & Codescu, M. (2019). *Hets User Guide, version 0.99*. Retrieved July 25, 2019 from http://www.informatik.unibremen.de/agbkb/forschung/formal_methods/CoFI/hets/

OCL. (2014). *OMG Object constraint language (OCL), version 2.4. OMG Document Number: formal/2014-02-03*. Retrieved July, 25, 2019 from <http://www.omg.org/spec/OCL/2.4>

OPEN FL. (2019). *OPEN FL 4.7*. Retrieved July 25, 2019 from <http://www.openfl.org/>

QVT. (2016). *QVT: MOF 2.0 query, view, transformation: Version 1.3*. OMG Document Number: formal/2016-06-03. Retrieved July 25, 2019 from <http://www.omg.org/spec/QVT/1.3>

Raibulet, C., Arcelli Fontana, F., & Zanoni, M. (2017). Model-Driven Reverse Engineering Approaches: A Systematic Literature Review. *IEEE Access: Practical Innovations, Open Solutions*, 5, 14516–14542. doi:10.1109/ACCESS.2017.2733518

Stankovic, J. (2014). Research Directions for the Internet of Things. *IEEE Internet of Things Journal*, 1(1), 3–9.

Steinberg, D., Budinsky, F., Paternostro, M., & Merks, E. (2009). *EMF: Eclipse Modeling Framework* (2nd ed.). Addison-Wesley.

Umuhoza, E., & Brambilla, M. (2016). Model Driven Development Approaches for Mobile Applications: A Survey. In *Proceedings of Mobile Web and Intelligent Information Systems - 13th International Conference, MobiWIS 2016* (pp. 93-107). Berlin: Springer. 10.1007/978-3-319-44215-0_8

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of Things for Smart Cities. *IEEE Internet of Things Journal*, 1(1), 22–32. doi:10.1109/JIOT.2014.2306328

KEY TERMS AND DEFINITIONS

ADM (Architecture Driven Modernization): The process of understanding and evolving existing software assets of an existing system in the context of MDA.

ATL (Atlas Transformation Language): A model transformation language and toolkit developed on top of the Eclipse platform that provides ways to produce target models from source models.

Ecore Metamodel: The de-facto reference implementation of EMOF (Essential Meta-Object Facility), a subset of MOF. It is the core metamodel of EMF.

Haxe: An open-source toolkit based on a modern, high level, strictly typed programming language, a cross-compiler, a complete cross-platform standard library and ways to access each platform's native capabilities.

KDM (Knowledge Discovery Metamodel): The core metamodel of ADM, a language-independent metamodel for representing assets of software legacy.

Metamodeling: The process of generating a “model of models”; the essence of Model Driven Development approaches.

Mobile Technology: The technology used for cellular communication.

Model Transformation: A mechanism for automatically creating target models based on information contained in existing source models

Model-Driven Architecture (MDA): An initiative of the OMG for the development of software systems based on the separation of business and application logic from underlying platform technologies. It is an evolving conceptual architecture to achieve cohesive model-driven technology specifications.

Model-Driven Engineering: Software engineering discipline that emphasizes the use of models and model transformations to raise the abstraction level and the degree of automation in software development.

Software Migration: A kind of modernization for moving from the use of one operating environment to another operating environment that is, in most cases, thought to be better.

Interaction Speed as Nonverbal Cues in Text Messaging via Smartphone

Kiminori Usuki

Tokyo Metropolitan Mizuho Noge High School, Japan

Shogo Kato

Tokyo Woman's Christian University, Japan

Yuuki Kato

 <https://orcid.org/0000-0002-2761-7735>

Sagami Women's University, Japan

INTRODUCTION

Computer-mediated communication (CMC) has the ability to bridge people across time zones and geographic regions (Kiesler & Sproull, 1992), thereby removing temporal and spatial barriers to interpersonal communication (Hesse, Werner, & Altman, 1988; Noman, 1983). Indeed, by using the Internet, people can communicate anytime, anywhere. Especially in text-based asynchronous forms of communication, such as email and electronic message boards, users can communicate at their convenience (Christie & de Alberdi, 1985; Katz & Aakhus, 2002; Tyler & Tang, 2003). However, this freedom from temporal restrictions—one of the two barriers relieved by CMC—has been weakened by the recent advent of mobile text-messaging (Kato, Kato, & Usuki, 2019; Murtagh, 2002). Expansion of communication devices from PCs to mobile phones and smartphones has greatly contributed to this (Kato & Kato, 2015). For example, people looking at their phones while walking is now a common sight on the street and in train stations (Lamberg & Muratori, 2012; Pešić, Antić, Glavić, & Milenković, 2016). Although they do so for a variety of reasons, it seems likely that many are texting while walking in order to quickly reply to received messages (Lim, Amado, Sheehan, & Van Emmerik, 2015; Schwebel, McClure, & Porter, 2017). In text messaging with mobile phones and smartphones, many users face pressure to rapidly reply (Kato et al., 2019; Kato & Kato, 2015, 2016). Despite the increasing importance of mobile text-messaging, few studies have examined reply speed. This article explores the roles played by reply speed in smartphone messaging, specifically focusing on messaging apps that include a read receipt function (i.e., a function that notifies the sender of whether a recipient has read a message) (Kato, Kato, & Ozawa, 2017). Additionally, this article discusses interaction speed as a form of nonverbal cue in text messaging.

BACKGROUND

Early studies of Internet-related psychology cited a lack of nonverbal cues and visual anonymity as characteristics of text-based communication on PCs (Joinson, 2003). A number of studies on CMC have discussed these characteristics of text-based CMC exchanges in terms of the absence of facial expressions, gestures, and nonverbal cues such as tone of voice that are involved in face-to-face conversation

DOI: 10.4018/978-1-7998-3479-3.ch062

(Kiesler, Siegel, & McGuire, 1984). These characteristics form the basis of various models and theories in CMC research conducted prior to the 1990s. For example, the social presence model (Short, Williams, & Christie, 1976) posited that when media conveys little nonverbal information, the sense of the other party as an actual person becomes less distinct—resulting in cold, impersonal communication. In media selection, media richness theory (Daft & Lengel, 1984; Daft & Lengel, 1986) recommended using rich media that can transmit information sufficient for unequivocal communication. The cuelessness model states that because CMC lacks the nonverbal cues that are present in face-to-face communication, communication tends to become task-oriented rather than socioemotional (Rutter, 1984; Rutter, 1987; Rutter, Stephenson, & Dewey, 1981). The reduced social cues approach says that in CMC lacking nonverbal information, users experience depersonalization, which results in uninhibited communications and consequently a higher probability of behaviors that are outside of social norms (McGuire, Kiesler, & Siegel, 1987; Siegel, Dubrovsky, Kiesler, & McGuire, 1986; Sproull & Kiesler, 1986).

An active area of CMC research in the 1980s was the exploration of the consequences of this deficiency. However, subsequent CMC studies have shown many examples of socioemotional communication that can deepen personal relationships, even in text message exchanges that contain few nonverbal cues (Walther, 2011). For instance, more self-disclosure is thought to occur in CMC (Joinson, 2001; Matheson, 1991; Matheson & Zanna, 1988). With few nonverbal cues, visual anonymity is preserved and de-individuation occurs, which attenuates the restraint typically exercised during face-to-face communication. Also, according to the hyper-personal model (Walther, 1996), the dearth of nonverbal cues in CMC permits interactions that are closer to the idealized self. For example, users can make controlled, selective presentations of themselves to the other party in a way that facilitates creating a favorable impression. These studies provided examples of the skillful employment of minimal nonverbal cues in CMC (Walther & Tidwell, 1995).

Research conducted over the last two decades reveals the importance of chronemic nonverbal cues in CMC, and how chronemics influence online interaction (Kalman, Scissors, Gill, & Gergle, 2013). These studies explored the influence of chronemic variables such as the time of day or length of the pause preceding a response on impression formation during online interactions (Döring & Pöschl, 2008; Kalman et al., 2013; Kalman & Rafaeli, 2011; Sheldon, Thomas-Hunt, & Proell, 2006). For example, Ballard and Seibold (2004) described how chronemics are a core component of human communication. Sheldon et al. (2006) focused on perceptions of delays in online replies and found that a delayed response by a high-status respondent was interpreted more positively than the same delay by a low-status respondent. Responsiveness is an important aspect of communication, and we are attuned to identify even slight delays in responsiveness. Such delays often lead to negative consequences, such as being perceived as less socially competent (McLaughlin & Cody, 1982). An experimental study revealed that the same task-related e-mail message was perceived as more assertive when sent at night than when sent during the day (Walther & Tidwell 1995). Short response times can be interpreted as nonverbal cues of interpersonal closeness, immediacy, care, presence, and even submissiveness (Döring & Pöschl, 2008). In summary, previous research on chronemics has been mainly confined to the context of users' impression formation during PC-based communication.

NONVERBAL CUES IN TEXT MESSAGING VIA MOBILE PHONE

Most early CMC research focused on Internet-mediated communication via PCs. However, many people now use mobile phones or smartphones to communicate over the Internet. Communication using such

devices differs from PC-based communication in several ways (Döring & Pöschl, 2008). For example, emoticons and emoji are used more frequently in text-based communication that occurs via mobile phones than in PC-based communication (Tossell, Kortum, Shepard, Barg-Walkow, Rahmati, & Zhong, 2012). Additionally, text messaging on mobile phones tends to be briefer and more interactive than PC-based communication (Mahatanankoon & O'Sullivan, 2008). The lack of nonverbal information in text-based communication has been the subject of much CMC research. However, now that mobile phones have come to play a major role in communication, response times can be thought of as conveying important nonverbal information (Kalman & Rafaeli, 2011). For example, a “slow” response might nonverbally convey negative emotion (Kalman et al., 2013). Acknowledging this potential negative cue, slow e-mail replies often contain apologies and explanations to compensate for the extended response time (Kalman & Rafaeli, 2011; Kalman, et al. 2013). In other words, the speed of interaction via PCs provides information that influences impression formation, such as of the user's individual characteristics, but the speed of mobile interaction provides more real-time information on emotional and psychological states.

In most text-based communication, including conventional email, there is no interaction while the message sender awaits a reply. However, messaging apps such as Facebook Messenger, WhatsApp, and LINE (an instant messenger application for smartphones, used mainly in Japan) have “read receipt” functions that notify senders when recipients have read a message, thereby allowing the sender to know that the recipient has read the message, even before any reply arrives (Hoyle, Das, Kapadia, Lee, & Vaniea, 2017; Kato et al., 2017). In such exchanges, many senders will be concerned not only about the time spent waiting for a reply, but also the time until the read receipt appears (Kato et al., 2017). One reason for this is that the sender may suspect that the recipient is ignoring the message if the read receipt is not displayed after an extended time (“ignored unread”) or if no reply is received long after the read receipt has been displayed (“ignored read”). Conversely, recipients may worry about their intentions being misunderstood if messages are left in an “ignored unread” state when messages cannot be read immediately, or in an “ignored read” state when unable to immediately reply to read messages. This psychological pressure is becoming problematic, especially among young users (Hoyle et al., 2017; Kato, Kato, & Ozawa, 2018; Kato & Kato, 2016).

From their presumed desire for a quick reply, it can be said that the sender is in a dominant position in smartphone-based text messaging (Kato et al., 2018). For example, message recipients may feel pressured to reply quickly so as to avoid causing anxiety in senders awaiting a response (Kato & Kato, 2015). However, reply speed is an operable nonverbal cue. Furthermore, as an addition to reply speed, text messaging with a read receipt function informs the sender whether a message has been read (Hoyle et al., 2017; Kato et al., 2017). This function, too, can be thought of as a nonverbal cue through which recipients can intentionally convey psychological meaning. Namely, in text messaging, where the sender's position dominates that of the recipient as described above, the reply speed and the read receipt function may provide strong nonverbal information that the recipient can manipulate. Strategic use of the read receipt is a nonverbal cue that enriches socioemotional interaction.

SOLUTIONS AND RECOMMENDATIONS

As the read receipt function continues to be adopted by additional smartphone messaging apps, our textual interactions are likely headed toward new temporal restrictions unlike those experienced during the initial spread of CMC. Restricted communication such as CMC offers individuals the ability to communicate more strategically (Walther & Parks, 2002). Kato et al. (2017) suggested that the emotional

states of individuals waiting for a reply can be affected by reply speed. Rapid exchanges of text messages in asynchronous media will eventually lead to frequent message exchanges and be positively related to the maintenance of socioemotional relationships (Walther, 1992). Therefore, quick replies in messaging serve strategic relational maintenance functions (Brody, Mooney, Westerman, & McDonald, 2009). However, it is possible to strategically manipulate reply speeds to elicit negative emotions in individuals waiting for a reply. For example, if a sender want to keep a romantic partner in anticipation through replying slowly, it might be more effective to let the partner wait with the message in read status than in unread status. This example is only one possibility, and strategic control of reply speed in messaging with a read receipt function is an interesting topic for future research. However, some studies make clear the possibility that reply speed in messaging with a read receipt display serves a role as a nonverbal cue that can affect senders' emotional states (Hoyle et al., 2017; Kato et al., 2017). These nonverbal cues are more difficult to interpret than are emoticons, so we must take care in attribution of meaning. Such cues may contain specific communicative intent in some cases, while in other cases there is no such intent—the delayed response is simply due to unavoidable circumstances that prevent a reply.

FUTURE RESEARCH DIRECTIONS

In this article, it was suggested that the speed of exchange and the display of read receipts in mobile text messaging have become new nonverbal cues. However, research has not yet evaluated strategies applied with these cues and their possible contribution to effective communication. Historically, various innovations have been made in text-based communication to alleviate constraints on expressing meaning, such as using emoticons composed of punctuation marks, letters, and symbols. Perhaps seasoned users of text messaging may already be using these new chronemic cues effectively. Therefore, insights may be gained via detailed analysis of seasoned users' behavior in terms of exchange speed and use of the read receipt function. In addition, it is necessary to verify the hypotheses concerning innovations obtained from observation and interview survey. Such data would assist researchers in formulating ideas aimed at modifying exchange speed and use of the read receipt function, which are now a source of psychological pressure for many users of text messaging applications.

Today, educational activities (e.g., Hossain, Nurunnabi, Hussain, & Saha, 2019; Hossain, Shan, & Nurunnabi, 2019) and businesses (e.g., Hossain, 2019) using mobile devices and smartphones have also become popular. In such situations, studying the effect of speed of interaction on effective learning and purchasing behavior is also a future issue. Additionally, previous research has pointed out that the recipient type, such as friendships, romantic relationships, and family relationships, influences demands for reply speed (Kato & Kato, 2015). In other words, it should be noted that the pressure for a quick reply varies depending on the recipient type. Finally, “read receipt functions” are functions added recently in mobile message texting, and this function speeds up the exchange of messages. The impact of new functions born through various upgrades on the speed of message exchange will change in the future. It will be necessary to keep an eye on the relationship between frequent updates on texting and chronemic nonverbal cues.

CONCLUSION

This article focused on the speed of interaction in text messaging via smart phone applications (especially those with read receipt functions). It was suggested that the speed of exchange and read receipt status can serve as operable nonverbal cues to the sender of the message. Toward the aim of effective communication via text messaging using smartphones, we propose future directions in examining the strategic use of speed of interaction and read receipts.

REFERENCES

- Ballard, D. I., & Seibold, D. R. (2004). Organizational members' communication and temporal experience: Scale development and validation. *Communication Research*, 31(2), 135–172. doi:10.1177/0093650203261504
- Brody, N., Mooney, C., Westerman, S., & McDonald, P. (2009). Its gt 2gthr l8r: Text messaging as a relational maintenance tool. *Kentucky Journal of Communication*, 28(2), 109–127.
- Christie, B., & de Alberdi, M. (1985). Electronic meetings. In B. Christie (Ed.), *Human factors of information technology in the office* (pp. 97–144). Wiley & Sons.
- Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organizational design. *Research in Organizational Behavior*, 6, 191–233.
- Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 32(5), 554–571. doi:10.1287/mnsc.32.5.554
- Döring, N., & Pöschl, S. (2008). Nonverbal cues in mobile phone text messages: The effects of chronemics and proxemics. In R. Ling & S. W. Campbell (Eds.), *The reconstruction of space and time: Mobile communication practices* (Vol. 109). Transaction Publishers.
- Hesse, B. W., Werner, C. M., & Altman, I. (1988). Temporal aspects of computer-mediated communication. *Computers in Human Behavior*, 4(2), 147–165. doi:10.1016/0747-5632(88)90023-4
- Hossain, S. F. (2019). Social Networking and Its Role in Media Entrepreneurship: Evaluating the Use of Mobile Phones in the Context of Online Shopping – A Review. *Journal of Media Management and Entrepreneurship*, 1(1), 73–86. doi:10.4018/JMME.2019010105
- Hossain, S. F., Nurunnabi, M., Hussain, K., Saha, S. K., & Wang, S. (2019). Effects of variety-seeking intention by mobile phone usage on university students' academic performance. *Information & Communications Technology in Education*, 6(1), 1574692. doi:10.1080/2331186X.2019.1574692
- Hossain, S. F., Shan, X., & Nurunnabi, M. (2019). Is M-Learning a Challenge?: Students Attitudes Toward the Sustainable Learning and Performance. *International Journal of e-Collaboration*, 15(1), 21–37. doi:10.4018/IJeC.2019010102
- Hoyle, R., Das, S., Kapadia, A., Lee, A. J., & Vaniea, K. (2017). Was my message read?: Privacy and signaling on Facebook Messenger. *CHI'17 Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3838–3842.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2), 177–192. doi:10.1002/ejsp.36

Joinson, A. N. (2003). *Understanding the psychology of internet behavior*. Palgrave Macmillan.

Kalman, Y. M., & Rafaeli, S. (2011). Online pauses and silence: Chronemic expectancy violations in written computer-mediated communication. *Communication Research*, 38(1), 54–69. doi:10.1177/0093650210378229

Kalman, Y. M., Scissors, L. E., Gill, A. J., & Gergle, D. (2013). Online chronemics convey social information. *Computers in Human Behavior*, 29(3), 1260–1269. doi:10.1016/j.chb.2012.12.036

Kato, S., Kato, Y., & Usuki, K. (2019). Associations between dependency on Line text messaging and occurrence of negative emotions in Line group chats. In M. Desjarlais (Ed.), *The Psychology and Dynamics Behind Social Media Interactions* (pp. 188–209). IGI Global.

Kato, Y., & Kato, S. (2015). Reply speed to mobile text messages among Japanese college students: When a quick reply is preferred and a late reply is acceptable. *Computers in Human Behavior*, 44, 209–219. doi:10.1016/j.chb.2014.11.047

Kato, Y., & Kato, S. (2016). Mobile phone use during class at a Japanese women's college. In M. N. Yildiz & J. Keengwe (Eds.), *Handbook of Research on Media Literacy in the Digital Age* (pp. 436–455). IGI Global. doi:10.4018/978-1-4666-9667-9.ch021

Kato, Y., Kato, S., & Ozawa, Y. (2017). Nobody read or reply your messages: Emotional responses among Japanese university students. *International Journal of Cyber Behavior, Psychology and Learning*, 7(4), 1–11. doi:10.4018/IJCBPL.2017100101

Kato, Y., Kato, S., & Ozawa, Y. (2018). Desired speed of reply during text-based communication via smartphones: A survey of young Japanese adults. In R. T. Gopalan (Ed.), *Intimacy and Developing Personal Relationships in the Virtual World* (pp. 64–83). IGI Global.

Katz, J., & Aakhus, M. (2002). *Perpetual contact: Mobile communication, private talk, public performance*. Cambridge University Press. doi:10.1017/CBO9780511489471

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer mediated communication. *The American Psychologist*, 39(10), 1123–1134. doi:10.1037/0003-066X.39.10.1123

Kiesler, S., & Sproull, L. (1992). Group decision making and communication technology. *Organizational Behavior and Human Decision Processes*, 52(1), 96–123. doi:10.1016/0749-5978(92)90047-B

Lamberg, E. M., & Muratori, L. M. (2012). Cell phones change the way we walk. *Gait & Posture*, 35(4), 688–690. doi:10.1016/j.gaitpost.2011.12.005 PMID:22226937

Lim, J., Amado, A., Sheehan, L., & Van Emmerik, R. E. A. (2015). Dual task interference during walking: The effects of texting on situational awareness and gait stability. *Gait & Posture*, 42(4), 466–471. doi:10.1016/j.gaitpost.2015.07.060 PMID:26298159

Mahatanankoon, P., & O'Sullivan, P. (2008). Attitude toward mobile text messaging: An expectancy-based perspective. *Journal of Computer-Mediated Communication*, 13(4), 973–992. doi:10.1111/j.1083-6101.2008.00427.x

- Matheson, K. (1991). Social cues in computer-mediated negotiations: Gender makes a difference. *Computers in Human Behavior*, 7(3), 137–145. doi:10.1016/0747-5632(91)90003-J
- Matheson, K., & Zanna, M. P. (1988). The impact of computer-mediated communication on self-awareness. *Computers in Human Behavior*, 4(3), 221–233. doi:10.1016/0747-5632(88)90015-5
- McGuire, T. W., Kiesler, S., & Siegel, J. (1987). Group and computer-mediated discussion effects in risk decision making. *Journal of Personality and Social Psychology*, 52(5), 917–930. doi:10.1037/0022-3514.52.5.917
- McLaughlin, M. L., & Cody, M. J. (1982). Awkward silences: Behavioral antecedents and consequences of the conversational lapse. *Human Communication Research*, 8(1), 299–316. doi:10.1111/j.1468-2958.1982.tb00669.x
- Murtagh, G. M. (2002). Seeing the “rules”: Preliminary observations of action, interaction and mobile phone use. In B. Brown (Ed.), *Wireless world: Social and international aspects of the mobile age* (pp. 81–91). Springer. doi:10.1007/978-1-4471-0665-4_6
- Noman, D. A. (1983). The computer always rings twice. *Psychology Today*, 17(10), 46–50.
- Pešić, D., Antić, B., Glavić, D., & Milenković, M. (2016). The effects of mobile phone use on pedestrian crossing behaviour at unsignalized intersections – Models for predicting unsafe pedestrians behaviour. *Safety Science*, 82, 1–8. doi:10.1016/j.ssci.2015.08.016
- Rutter, D. R. (1984). *Looking and seeing the role of visual communication in social interaction*. John Wiley & Sons.
- Rutter, D. R. (1987). *Communicating by telephone*. Pergamon Press.
- Rutter, D. R., Stephenson, G. M., & Dewey, M. E. (1981). Visual communication and the content and style of conversation. *British Journal of Social Psychology*, 20(1), 41–52. doi:10.1111/j.2044-8309.1981.tb00472.x PMID:7237005
- Schwebel, D. C., McClure, L. A., & Porter, B. E. (2017). Experiential exposure to texting and walking in virtual reality: A randomized trial to reduce distracted pedestrian behavior. *Accident; Analysis and Prevention*, 102, 116–122. doi:10.1016/j.aap.2017.02.026 PMID:28279843
- Sheldon, O. J., Thomas-Hunt, M. C., & Proell, C. A. (2006). When timeliness matters: The effect of status on reactions to perceived time delay within distributed collaboration. *The Journal of Applied Psychology*, 91(6), 1385–1395. doi:10.1037/0021-9010.91.6.1385 PMID:17100492
- Short, J. A., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. Wiley.
- Siegel, J., Dubrovsky, V., Kiesler, S., & McGuire, T. W. (1986). Group processes in computer-mediated communication. *Organizational Behavior and Human Decision Processes*, 37(2), 157–187. doi:10.1016/0749-5978(86)90050-6
- Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science*, 32(11), 1492–1512. doi:10.1287/mnsc.32.11.1492

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H., Rahmati, A., & Zhong, L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior*, 28(2), 659–663. doi:10.1016/j.chb.2011.11.012

Tyler, J. R., & Tang, J. C. (2003). When can I expect an email response? A study of rhythms in email usage. *Proceedings of the Eighth European Conference on Computer Supported Cooperative Work (ECSCW2003)*, 239–258. 10.1007/978-94-010-0068-0_13

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research*, 19(1), 52–90. doi:10.1177/009365092019001003

Walther, J. B. (1996). Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1), 3–43. doi:10.1177/009365096023001001

Walther, J. B. (2011). Theories of computer-mediated communication and interpersonal relations. In M. L. Knapp & J. A. Daly (Eds.), *The SAGE handbook of interpersonal communication* (4th ed., pp. 443–479). Sage.

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in: Computer mediated communication and relationships. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of interpersonal communication* (Vol. 3, pp. 529–563). Sage.

Walther, J. B., & Tidwell, L. C. (1995). Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing and Electronic Commerce*, 5(4), 355–378. doi:10.1080/10919399509540258

ADDITIONAL READING

Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The role of emoticon in computer-mediated communication: A review. *Computers in Human Behavior*, 24(8), 766–785. doi:10.1016/j.chb.2007.04.004

Joinson, A. N., McKenna, K., Postmes, T., & Reips, U. D. (Eds.). (2007). *The oxford handbook of Internet psychology*. Oxford University Press.

Kato, Y., Kato, S., & Akahori, K. (2007). Effects of emotional cues transmitted in e-mail communication on the emotions experienced by senders and receivers. *Computers in Human Behavior*, 23(4), 1894–1905. doi:10.1016/j.chb.2005.11.005

Kaye, L. K., Malone, S. A., & Wall, H. J. (2017). Emojis: Insights, affordances, and possibilities for psychological science. *Trends in Cognitive Sciences*, 21(2), 66–68. doi:10.1016/j.tics.2016.10.007 PMID:28107838

Scott, D. J., Coursaris, C. K., Kato, Y., & Kato, S. (2009). The exchange of emotional context in business communications: A comparison of PC and mobile email users. In M. M. Head & E. Li (Eds.), *Mobile and ubiquitous commerce: Advanced e-business methods: volume 4 of advances in electronic business series* (pp.201–219). Hershey, PA: IGI Global.

Thompson, P. A., & Fougler, D. A. (1996). Effects of pictographs and quoting on flaming in electronic mail. *Computers in Human Behavior*, 12(2), 225–243. doi:10.1016/0747-5632(96)00004-0

Walther, J. B., & Burgoon, J. K. (1992). Relational communication in computer-mediated interaction. *Human Communication Research*, 19(1), 50–88. doi:10.1111/j.1468-2958.1992.tb00295.x

Walther, J. B., & D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19(3), 324–347. doi:10.1177/089443930101900307

KEY TERMS AND DEFINITIONS

Digital Immigrants: Digital immigrants are the generations preceding digital natives, who initially used traditional media before gradually adopting new ICTs and services as they became available.

Digital Natives: Digital natives are born into an information/communications technology environment where the Internet, email, and mobile phones are commonly used.

LINE Application: LINE is an instant messenger that is similar to other internet-based messengers, such as WhatsApp, and is the most-downloaded social media/messenger application in Japan. The application was released in June 2011, following the Great East Japan Earthquake of March 2011, with the aim of creating an application where users can easily communicate even in the event of a disaster. In Japan, LINE has come to replace traditional mobile text messaging services (including SMS), especially among young people.

Read Receipt Status: A read receipt function automatically informs the sender that the recipient has read the message.

Speed of Reply: Although text messaging can easily exchange text messages, its users typically expect faster replies than with cell phone text messaging. In fact, the act of periodically checking one's smartphone for incoming messages happens daily.

Stickers: The sticker is a newer graphical element similar to emoji or emoticons that conveys emotional state, attitude, and opinion and is a small illustration that is sent in place of a message. The sticker originated as an attachment in LINE, but by 2013, Facebook Messenger was equipped with similar features, followed by Facebook timeline in 2014. In 2016, iMessage for iPhone added sticker functionality.

Unread Receipt Status: An unread receipt function automatically informs the sender that the recipient has not read the message.

Narrowband Internet of Things

7

Sudhir K. Routray

 <https://orcid.org/0000-0002-2240-9945>

Bule Hora University, Bule Hora, Ethiopia

INTRODUCTION

Internet of things (IoT) is pervasive extension of the Internet. It uses several sensors based on the application requirements for interconnecting them together in a common framework. These days IoT has several dimensions for numerous applications and it facilitates massive machine type communications. IoT has the ability to reach the territories where there is no Internet available. Internet is normally available where there the human settlements are present. However, there are many scenarios where we need the information without the presence of the Internet. For instance, measurements of under surface soil humidity and fertilizer intensity can be brought in to the Internet through the IoT. This can be done by deploying appropriate sensors under the earth surface. Similarly, there are applications in which micro monitoring is required in the unmanned territories. In those cases, the supervision of the IoT can be done through some remote communication nodes. IoT can provide the information from many unmanned locations using its sensors and the communication infrastructure of the Internet.

Background of NBIoT

IoT is now an integral part of the modern digital ecosystem and leading the changing trends in many sectors. It has several ramifications and can be applied to a number of sensor-based application scenarios. Initially, it was started as a value-added segment for the cellular communication networks. Gradually, they became very popular in many applications. Now they appear in several cellular and non-cellular forms including sensor networks, control networks, and independent networks. Different types of IoTs are available depending on their features. Narrowband IoT (NBIoT) is one of the standardized forms of IoT which is very attractive due to its suitable features. According to the name, NBIoT needs narrow band frequencies for its operations. This narrow band requirement is quite special for NBIoT and thus it has several advantages over other forms of IoTs. NBIoT has ability to cover a large area with a small amount of transmit power (Routray & Sharmila, 2017). This is an energy efficient feature of NBIoT which makes it a popular green technology. These days green and sustainable technologies have high demand due to their energy efficiency and environment friendly features (Zhu et al., 2015). In these quest, NBIoT like low energy technologies are very popular in different applications in which IoTs and sensor-based technologies are essential. Now green IoTs are well researched areas in which optimization of resources is a prime goal (Routray & Sharmila, 2017). In the resource limited situations, such as the developing countries, they have even more important roles due to the shortage of resources.

In the recent years, several works have been done on different types of IoTs to make them more energy efficient. In Routray & Sharmila (2017), green IoTs have been studied from several technological aspects. In that work NBIoT has been proposed as the inherent green technology for several applications due to its attractive energy and bandwidth saving features. In Zhu et al. (2015), energy efficient and sustainable

DOI: 10.4018/978-1-7998-3479-3.ch063

initiatives for IoT based smart world have been addressed. It provides several recent progressive trends on smart initiatives and technologies which are able to make the entire IoT networks energy efficient. It also gives insight for the segment-wise greening processes in the framework of existing information and communication technologies (ICTs) across different application domains. Perera et al. (2014), provided a comprehensive survey of the contemporary IoT marketplace of last few years. That survey explains the recent industrial demands and applications of IoT and its associated technologies in different industrial and other associated processes. They also elaborate through justified logic that manufacturing and automation will have a large share of industrial IoT. Da Xu, He, Li (2014) too provided a similar survey which focuses on the technological prospective of IoT. This survey emphasizes that the wide deployments need to be energy efficient for the practical reasons. In Al-Fuqaha et al. (2015), a comprehensive survey on the practical aspects of IoT such as the enabling technologies, protocols and applications have been presented. At first it deals with the enabling technologies that make IoT a reality and then show the main protocols which are instrumental in the IoT operations. Finally, it provides a long list of applications in which IoT can play significant roles. The authors pointed that the overall success of IoT is very much dependent on the efficiency of the entire IoT networks, their components and end devices. Ramnath et al. (2017) and Yu et al. (2016), provide the hybridization of IoT networks with other types of networks, such as the cellular networks for advanced applications like the localization and tracking of objects and living animals. In the later one the 5G IoT related localization has been emphasized using special algorithms.

The work of Mohanty and Routray (2017) presents the recent trends of consumer electronics driven data communication networks. This paper shows the volume of the resources, such as the bandwidth and power for transmission needed for the global communication networks. The data storage and communication infrastructure required for the current ICT sectors too have been presented in this article. Sustainability aspects of communication networks have been presented in Mohanty & Moreira (2014). In this paper, carbon emission contributions of global communication sector have been presented which shows that more than 2% global carbon emissions are from the ICT sector. Outlooks for overall sustainability in the ICT sector have been proposed in this work. Pirinen (2014), provide the research and development initiatives of 5G and IoT are presented in which energy efficiency and overall network optimizations are emphasized. Hossain et. al. (2012), present the fundamental aspects of green radio communication technologies and the emerging methods used in ICT sectors for energy efficiency. This work also provides insights for emerging networking technologies such as IoT for overall energy efficiency. Cavalcante et al. (2014), present 5G energy efficiency related aspects which mentions general system characteristics for overall effectiveness on the system. Anand & Routray (2017), present the applications scenarios of NB-IoT for healthcare. This work analyses several issues and challenges in the healthcare NB-IoT. Security aspects of IoT are quite challenging. In Routray et al. (2017) novel cryptographic techniques for IoT have been suggested. Quantum cryptography for IoT is a robust security for the long term. Standardization NB-IoT was started in LTE Release 13 (Ratilainen, 2016). Basic specifications of NB-IoT were established for the cellular NB-IoT. These standards were updated for enhanced performances in Release 14 (Hoglund et al., 2017).

In this article, we provide the main principles, potentials and applications of NB-IoT. We present its deployment issues and the commonly encountered challenges in its implementation. These days NB-IoT is very popular in several situations due to its attractive features. We show that NB-IoT is the right choice for the developing countries at the moment. We also address its main motivating features, deployment options, and applications in different sectors.

The remainder of this article is organized in three sections. In Section 2, we present the energy efficient and green characteristics of NB-IoT. We also presented its deployment options and applications of

NB-IoT in different sectors. In Section 3, we present the overall initiatives being taken in the recent times to make NB-IoT resource efficient. We showed its suitability for the developing countries. In Section 4, we conclude with the main points of this article.

General Features of NB-IoT

NB-IoT is the low energy and low bandwidth version of IoT which is designed for the massive machine-to-machine communications. As the name suggests, it uses narrow bands for its different functions and operations. It needs a bandwidth of just 180 kHz to 200 kHz for its designated processes. NB-IoT finds a special status in the recent standards of Third Generation Partnership Project (3GPP). In their long term evolution (LTE) Release 13 and 14, 180 kHz has been proposed as the operating bandwidth for NB-IoT. It is a low power wide area (LPWA) technology which can save a lot of power when compared with other forms of IoTs. It is good for large scale economical deployment of IoT for different applications. In true sense, it is leaner, thinner, and greener than other IoTs proposed in the recent years. It can be deployed in both the cellular and non-cellular forms. However, cellular forms are popular as they can use the existing cellular architectures for its operations. In LTE Release 13 and 14, it has been standardized according to the compatible LTE provisions and also proposed for connected living environments (Hoglund et al., 2017).

Basic LPWA Features of NB-IoT

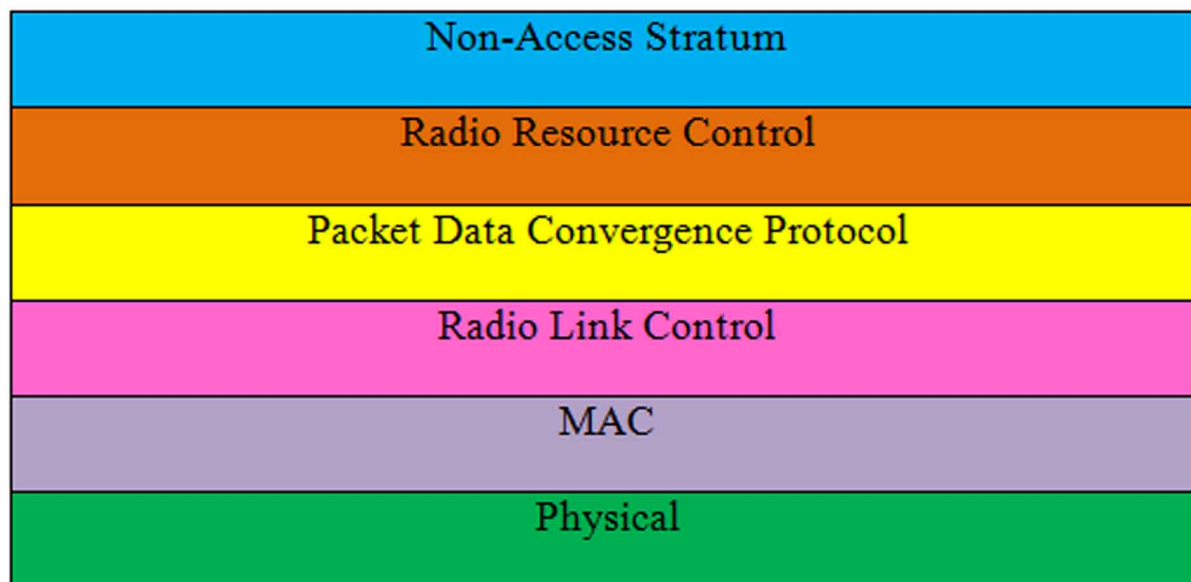
NB-IoT is an energy efficient version of IoT. The main attraction of NB-IoT is its LPWA nature. It can save a large amount of energy and bandwidth of a network. Its architecture and protocols have been standardized in the recent releases of LTE for different environments. Deployment bandwidths of NB-IoT have been agreed by the standardization committee in Release 13. Depending on the situation, different spectral bands can be used by different operators. According to Release 13 of 3GPP LTE, the maximum usable bandwidth of an end device is 200 kHz. Generally, for the communication purpose 180 kHz is used (Routray & Sharmila, 2017). At this bandwidth, the upper limits of the uplink and downlink data rates are set at 150 kbps. In Release 13, half duplex mode has been recommended for NB-IoT communications. In Release 14, these specifications have been further enhanced for advanced applications. The power from the transmitter is kept quite small so that a single battery can supply power to the NB-IoT devices or nodes for more than 10 years. In fact, in the recent NB-IoT standards, two power levels have been specified: 20 dBm and 23 dBm (Routray & Sharmila, 2017). The sensitivities of the NB-IoT sensors are really good. They can receive the signal at a power level as low as -64 dBm. These are the exclusive features of NB-IoT which are not found in other versions of IoTs proposed so far.

NB-IoT Architecture

A systematic NB-IoT architecture is required for its planning, dimensioning, cost estimation, design and final deployment. NB-IoT does not have a legacy to follow as it is one of the earliest IoTs of its type. However, it is similar to the wireless sensor networks (WSNs) to some extent and the WSNs have been there for several years. The existing WSN architecture and topologies can be helpful in the design of NB-IoT and its further advancement. It is noteworthy that WSNs did not have a structured and well-defined architecture like the cellular systems, such as the LTE networks which will form the backbone of NB-IoT. Therefore, an LTE cellular framework for NB-IoT is the right choice at the moment (Hoglund

et al., 2017). The layered architecture of NB-IoT is helpful in its planning and deployment. In Release 13, several specifications for different layers have been mentioned. NB-IoT can be separated into six layers as shown in Figure 1. The physical layer is at the bottom and it is normally the air interface. Physical layer does the similar functions as in WSNs and some additional functions as defined in Release 13. Above it is the medium access control (MAC) layer. This has the similar functions like the MAC layer of other networks. It incorporates the protocols for medium access and multiple access techniques. There is a radio link control layer in between the MAC and the upper layers. This layer makes the adaptation of the MAC layer information for radio links. Above it is the packet data convergence protocol layer which provides routing, traffic scheduling, networking, and other related tasks. Then above it is the radio resource control layer which takes care of the radio resources of the packets in the channels. NB-IoT uses user datagram protocol (UDP) and other cellular mechanism to carry this function. UDP is effective in the wireless networks and thus suitable for NB-IoT as well. The topmost layer is the Non-Access Stratum (NAS) which establishes the communication between the user equipment (UE) and the main server of the NB-IoT, also known as NB-IoT central node. NAS also provides the security and privacy related operations to the UE.

Figure 1. OSI Model of NB-IoT with six layers architecture which follows the LTE design to a large extent

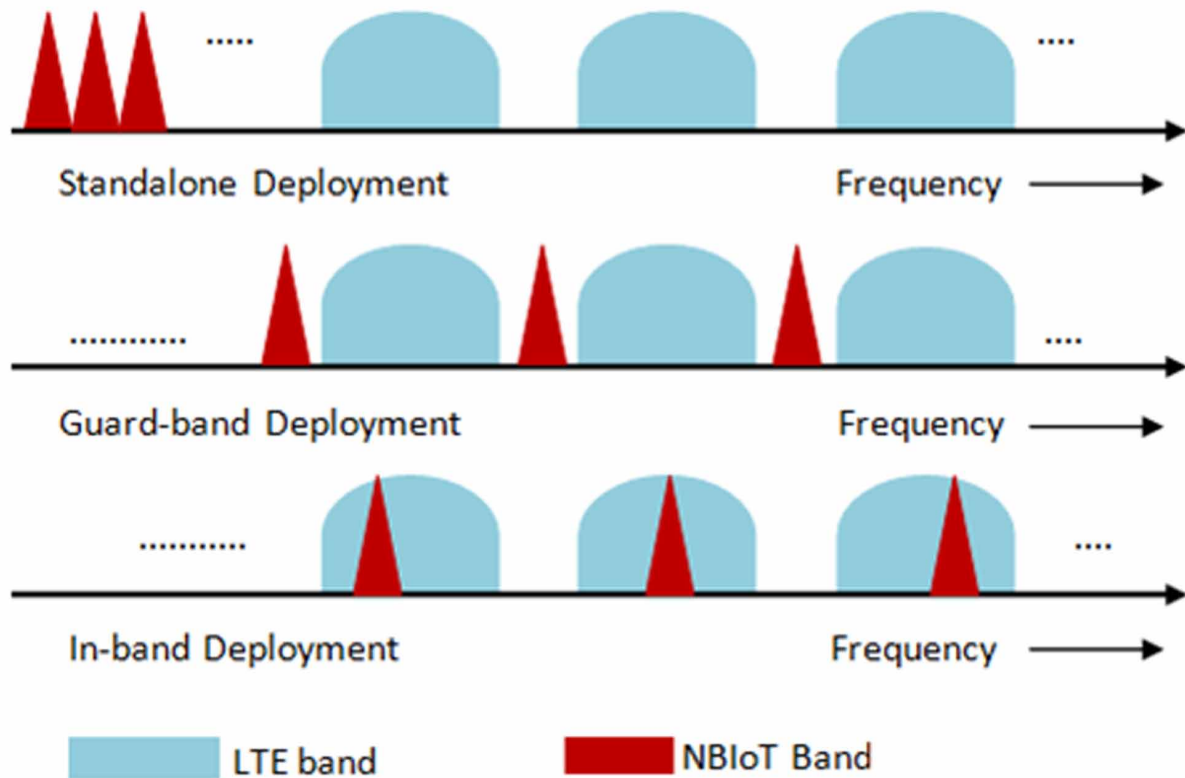


Deployment Options for NB-IoT

NB-IoT deployment options have been agreed in LTE Release 13. According to it, NB-IoT can be deployed in three different forms. The first being the standalone deployment in which an independent microwave band is provided in the 700MHz or 800 MHz range for its deployment. The second one is the guard band deployment in which the spectrum used for NB-IoT is the unused GSM and LTE guard bands. These guard bands are normally not used for communication purposes in GSM and LTE cellular systems. With the arrival of NB-IoT, these guard bands find a new application. The third one is the in-band deployment. In this case, some part of the GSM or LTE spectrum is allocated to the NB-IoT. In Fig. 2, we show these 3

deployment scenarios. In practice, either of them can be used depending on the availability of the bands and its suitability. If required, even two or all three options can also be used in specific cases.

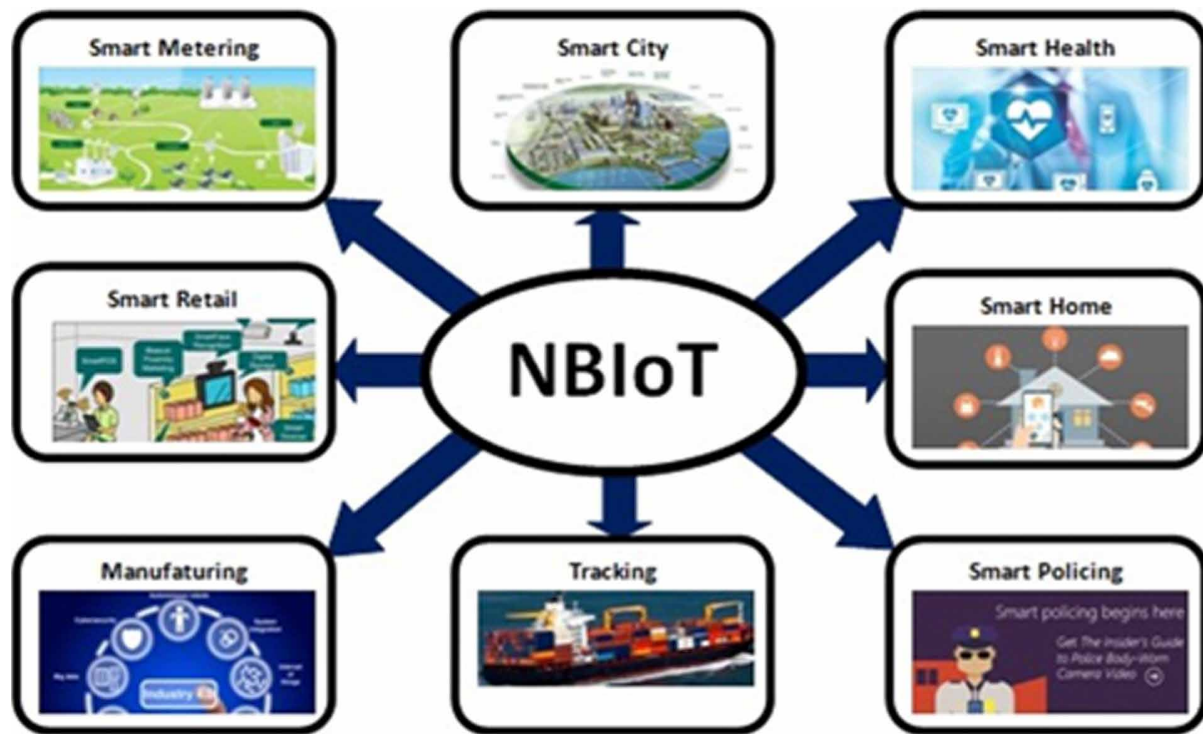
Figure 2. NBIoT deployment options in different bands



Applications of NBIoT

NBIoT is attractive in the LPWA regime. It has tremendous potential to operate in the resource limited scenarios. Here, we shed the light on some of the typical sectors where it is currently considered for deployment. Several applications of NBIoT are possible and it can be deployed for any LPWA applications in which the all the relevant elements of IoT are available. Both local and nonlocal applications of NBIoT are equally possible. Emerging applications are found in several technology and social sectors with every passing month. There are several new sectors in which NBIoT has potential applications. Some of these applications are indoor and the others are outdoor. NBIoT can be instrumental in designing smart homes and smart cities (Routray & Sharmila, 2017). Similarly, it can help in healthcare as well as the security monitoring of public places. In Fig. 3, we present some typical sectors in which NBIoT can have significant role. These sectors are: smart city management and monitoring; smart home applications; large scale manufacturing; pet tracking; kids' tracking; healthcare monitoring; safety and security related applications; smart agriculture and farming; energy and utility management; automobile and vehicular management; retail management; policing and law enforcement assistance; detection of environmental degradation; waste management; and several other LPWA applications. In fact, for rural areas NBIoT is considered the only choice due to its economical features.

Figure 3. Different applications of NBIoT. New applications emerge every month in different sectors.



NBIoT for Agriculture

NBIoT can be deployed in the agricultural sector for several performance enhancement related aspects. In agriculture, several resources are used to increase the harvest. Many of those resources are under-utilized due to the lack of facilities to monitor their utilities. For optimal use of the resources, NBIoT can be used which can sense the level of utilization and improve the utilization efficiency. For instance, water is an essential resource for agriculture. It is not available in sufficient amount in every part of the world. In the dry areas, water has to be utilized very efficiently. This utilization can be monitored in the agricultural farms through the water sensors deployed under the soil. Similarly, fertilizer concentration, pesticide intensity, and humidity levels too can be monitored using different types of sensors as a part of NBIoT in agriculture. These resources can also be remotely controlled through the NBIoT. As the energy levels for these sensing operations are very low, the cost of this IoT based monitoring is very much economical for the farmers. On the other hand, the harvest can be maximized due to the efficient nutrition distribution to the plants. These initiatives are already being deployed in several countries and found to be very effective. Of course, it will be made even more efficient in the future.

NBIoT for Manufacturing

Manufacturing is an important economic sector for developing countries. Recently, manufacturing companies in the developing countries are experiencing extreme competition, mainly due to the increasing pressures from technological changes and global business challenges. Advanced manufacturing technologies strongly rely on various ICT technologies to achieve higher productivity, higher quality,

and lower production costs. However, the role of the Internet in manufacturing is still understudied. IoT is going to fill this void very fast. NBIoT is even more efficient to handle the issues of automation and energy efficiency in large manufacturing hubs (Routray & Sharmila, 2017). Thus, it is a hot choice for the industrial automation in the developing countries. In all these changing scenarios, overall efficiency and automation can be provided by NBIoT (Routray & Sharmila, 2017).

NBIoT for Healthcare

Healthcare provisioning in any country is one of the basic services. It is also a large sector due to the very nature of the services provided by this system. Healthcare system is multi-tiered and need proper collaboration among each of the tiers. In several healthcare processes, NBIoT can play a role of game changer. For instance, in emergency proceedings it can help the patients to get appropriate treatments while coming to the hospital from their residence. Using the sensors, the vital information of the patients can be sent to the doctor who in return can provide the appropriate support through the sensors till the patient reaches the hospital. Similarly, the telemedicine applications too can be improved through the NBIoT based sensor assistance.

NBIoT for Resource Management

Basic resources such as water, electricity, gas and other essential commodities are not properly managed in the developing countries. That leads to the poor resource utilization efficiency and loss of revenue for the government and associated public sectors. NBIoT based solutions already exist to improve the resource utilization efficiency in these areas. Smart metering and efficient resource distribution can be a game changer in these sectors. NBIoT based automation in the metering and resource distribution is quite cost effective and efficient. It can provide a real-time resource monitoring framework to the resource distributors. This is how the operators can directly monitor the efficiency of resource utilization. Thus, NBIoT is being deployed in several countries for the public resource management.

Extensions of NBIoT

NBIoT can be extended for several new services and applications. For instance, it can be extended for undersea and space applications. Undersea applications are measurement of ocean parameters such as pressure, temperature, salinity and tidal velocity; location finding and tracking of objects on the sea surface. Similarly, the marine environment of the sea surface can be monitored using NBIoT sensors. Satellite integration with NBIoT can provide a lot of new applications. Satellites (in fact, the satellite constellation) provide better coverage of services than the cellular networks. Reliability of satellite links is better than the cellular links and thus dependable for critical applications. Satellites with a well-managed constellation provide more than 99.9% availability which is much better than the cellular networks. In mission critical applications such as disaster management and real-time rescue operations satellite based IoTs are preferable over the cellular ones. More about these extensions have been presented in the following section.

Future Research Directions

NB-IoT is new in the cellular networks. However, due to its attractive features it becomes very popular in the LPWA regime. It was first standardized in LTE release 13 which was its first standardization for large scale deployment. In Release 14 several enhancements have been proposed for new applications. In the future there will be many more additions for its better function. The new directions considered for NB-IoT are: satellite NB-IoT integration for better coverage and wider reach; location determination; implementation of mission critical applications through NB-IoT; monitoring and measurement of physical parameters at difficult locations/ places. Satellite integration with NB-IoT will be a great advantage for its commercial applications. Cellular networks are not very much reliable all the time. The availability of the cellular networks these days is less than 99.9%. In the rural areas it is even lower. It can be enhanced to 99.99% or even better through the satellites (Routray & Hussein, 2018).

Location determination using NB-IoT is quite interesting. Location determination using the cellular networks has limited accuracy due to the sparse presence of the base stations. However, NB-IoT uses many sensors under each base station and it can provide much better accuracy than the cellular networks (Jiang et al., 2013). Due to the better accuracy in the locations, tracking too is very effective using NB-IoT (Ramnath et al., 2017). Location based services will be very much dependent on NB-IoT like infrastructure. These are the future value added services of NB-IoT and currently being researched. Location determination through the IoTs can be extended to both indoor and outdoor environments (Gu & Ren, 2015).

Disaster management and several other emergency services are very much dependent on the locations. For these services time is a critical resource due to its limitation. On the other hand these services must be provided at the right time with successful execution. Thus these services are known as mission critical applications. NB-IoT is a good choice for mission critical applications due to its widespread coverage. Satellite integrated NB-IoT can be very effective for these mission critical applications (Routray & Hussein, 2018).

There are several locations or places where human presence for monitoring and measurement are dangerous. Such cases can be handled through remotely placed sensors or sensor networks. NB-IoT is a low power and highly efficient sensor network. Therefore, NB-IoT is being researched for the deployment in such difficult places. For instance, ocean surfaces and floors needs to be monitored and its parameters such as the temperature, pressure and salinity needs to be measures regularly. This can be done through NB-IoT. For these applications, satellite coordination can be used. Even wired networks along with the NB-IoT sensors can execute these tasks efficiently. NB-IoT has enormous potential in the practical applications. Therefore, the future research directions associated with NB-IoT will have several ramifications.

CONCLUSION

In this article, the main advantages and attractive features of NB-IoT have been highlighted. NB-IoT is here to stay for a long time. It provides several advantages which are not found in other IoTs. It can be considered as a green IoT due to its energy efficiency and narrow bandwidth requirements. Main design related features of NB-IoT are borrowed from LTE. Thus the design methodologies of NB-IoT follow the LTE like framework. It is also true that NB-IoT is compatible with almost all types of the legacy networks of LTE. Its potentials are far reaching. NB-IoT is an LPWA technology that makes it more attractive over

other IoTs. This is especially important for the developing countries due to its economical deployment over other forms of IoTs.

REFERENCES

- Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys and Tutorials*, 17(4), 2347–2376. doi:10.1109/COMST.2015.2444095
- Anand, S., & Routray, S. K. (2017). Issues and challenges in healthcare narrowband IoT. In *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on* (pp. 486–489). IEEE. 10.1109/ICICCT.2017.7975247
- Anand, S., & Routray, S. K. (2017, March). Issues and challenges in healthcare narrowband IoT. In *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on* (pp. 486–489). IEEE. 10.1109/ICICCT.2017.7975247
- Cavalcante, R. L., Stanczak, S., Schubert, M., Eisenblaetter, A., & Tuerke, U. (2014). Toward energy-efficient 5G wireless communications technologies: Tools for decoupling the scaling of networks from the growth of operating power. *IEEE Signal Processing Magazine*, 31(6), 24–34. doi:10.1109/MSP.2014.2335093
- Chen, Z., Xia, F., Huang, T., Bu, F., & Wang, H. (2013). A localization method for the Internet of Things. *The Journal of Supercomputing*, 63(3), 657–674. doi:10.1007/11227-011-0693-2
- Colombo, A., Fontanelli, D., Macii, D., & Palopoli, L. (2014). Flexible indoor localization and tracking based on a wearable platform and sensor data fusion. *IEEE Transactions on Instrumentation and Measurement*, 63(4), 864–876. doi:10.1109/TIM.2013.2283546
- Da Xu, L., He, W., & Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4), 2233–2243. doi:10.1109/TII.2014.2300753
- Gu, Y., & Ren, F. (2015). Energy-efficient indoor localization of smart hand-held devices using Bluetooth. *IEEE Access: Practical Innovations, Open Solutions*, 3, 1450–1461. doi:10.1109/ACCESS.2015.2441694
- Hoglund, A., Lin, X., Liberg, O., Behravan, A., Yavuz, E. A., Van Der Zee, M., Sui, Y., Tirronen, T., Ratilainen, A., & Eriksson, D. (2017). Overview of 3GPP release 14 enhanced NB-IoT. *IEEE Network*, 31(6), 16–22. doi:10.1109/MNET.2017.1700082
- Hossain, E., Bhargava, V. K., & Fettweis, G. P. (Eds.). (2012). *Green radio communication networks*. Cambridge University Press. doi:10.1017/CBO9781139084284
- Jiang, J. A., Zheng, X. Y., Chen, Y. F., Wang, C. H., Chen, P. T., Chuang, C. L., & Chen, C. P. (2013). A distributed RSS-based localization using a dynamic circle expanding mechanism. *IEEE Sensors Journal*, 13(10), 3754–3766. doi:10.1109/JSEN.2013.2258905
- Mohanty, S., & Moreira, A. C. (2014). Sustainability in global telecommunications. *IEEE Potentials*, 33(5), 29–34. doi:10.1109/MPOT.2013.2292540

Mohanty, S., & Routray, S. K. (2017). CE-Driven Trends in Global Communications: Strategic sectors for economic growth and development. *IEEE Consumer Electronics Magazine*, 6(1), 61–65. doi:10.1109/MCE.2016.2614420

Perera, C., Liu, C. H., Jayawardena, S., & Chen, M. (2014). A survey on internet of things from industrial market perspective. *IEEE Access: Practical Innovations, Open Solutions*, 2, 1660–1679. doi:10.1109/ACCESS.2015.2389854

Petcovici, A., & Stroulia, E. (2016, December). Location-based services on a smart campus: A system and a study. In *Internet of Things (WF-IoT), 2016 IEEE 3rd World Forum on* (pp. 94–99). IEEE. 10.1109/WF-IoT.2016.7845406

Pirinen, P. (2014, November). A brief overview of 5G research activities. In *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on* (pp. 17–22). IEEE.

Ramnath, S., Javali, A., Narang, B., Mishra, P., & Routray, S. K. (2017, May). IoT based localization and tracking. In *IoT and Application (ICIOT), 2017 International Conference on* (pp. 1–4). IEEE. 10.1109/ICIOTA.2017.8073629

Ratilainen, A. (2016). *NB-IoT presentation for IETF LPWAN*. IETF. Available: <https://www.ietf.org/proceedings/97/slides/slides-97-lpwan-30-nbiot-presentation-00.pdf>

Routray, S. K., & Hussein, H. M. (2018). Satellite Based IoT Networks for Emerging Applications. In *Advances Wireless Africa, 2018 First International Conference on* (pp. 1–5). IEEE.

Routray, S. K., Jha, M. K., Sharma, L., Nyamangoudar, R., Javali, A., & Sarkar, S. (2017, May). Quantum cryptography for IoT: A Perspective. In *IoT and Application (ICIOT), 2017 International Conference on* (pp. 1–4). IEEE. 10.1109/ICIOTA.2017.8073638

Routray, S. K., & Sharmila, K. P. (2017). Green initiatives in IoT. In *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on* (pp. 454–457). IEEE. 10.1109/AEEICB.2017.7972353

Wang, C., Daneshmand, M., Dohler, M., Mao, X., Hu, R. Q., & Wang, H. (2013). Guest Editorial-Special issue on internet of things (IoT): Architecture, protocols and services. *IEEE Sensors Journal*, 13(10), 3505–3510. doi:10.1109/JSEN.2013.2274906

Wang, Y. P. E., Lin, X., Adhikary, A., Grovlen, A., Sui, Y., Blankenship, Y., Bergman, J., & Razaghi, H. S. (2017). A primer on 3GPP narrowband Internet of Things. *IEEE Communications Magazine*, 55(3), 117–123. doi:10.1109/MCOM.2017.1600510CM

Xia, F., Yang, L. T., Wang, L., & Vinel, A. (2012). Internet of things. *International Journal of Communication Systems*, 25(9), 1101–1102. doi:10.1002/dac.2417

Yu, R., Bai, Z., Yang, L., Wang, P., Move, O. A., & Liu, Y. (2016). A Location Cloaking Algorithm Based on Combinatorial Optimization for Location-Based Services in 5G Networks. *IEEE Access: Practical Innovations, Open Solutions*, 4, 6515–6527. doi:10.1109/ACCESS.2016.2607766

Zhu, C., Leung, V. C., Shu, L., & Ngai, E. C. H. (2015). Green internet of things for smart world. *IEEE Access: Practical Innovations, Open Solutions*, 3, 2151–2162. doi:10.1109/ACCESS.2015.2497312

ADDITIONAL READING

Li, S., Da Xu, L., & Zhao, S. (2018). 5G internet of things: A survey. *Journal of Industrial Information Integration*.

Li, Y., Cheng, X., Cao, Y., Wang, D., & Yang, L. (2018). Smart choice for the smart grid: Narrow-band Internet of Things (NB-IoT). *IEEE Internet of Things Journal*, 5(3), 1505–1515. doi:10.1109/JIOT.2017.2781251

Petrov, V., Samuylov, A., Begishev, V., Moltchanov, D., Andreev, S., Samouylov, K., & Koucheryavy, Y. (2017). Vehicle-based relay assistance for opportunistic crowdsensing over narrowband IoT (NB-IoT). *IEEE Internet of Things journal*.

Xu, T., & Darwazeh, I. (2018). Non-Orthogonal Narrowband Internet of Things: A Design for Saving Bandwidth and Doubling the Number of Connected Devices. *IEEE Internet of Things Journal*, 5(3), 2120–2129. doi:10.1109/JIOT.2018.2825098

Yang, X., Wang, X., Wu, Y., Qian, L. P., Lu, W., & Zhou, H. (2018). Small-Cell Assisted Secure Traffic Offloading for Narrowband Internet of Thing (NB-IoT) Systems. *IEEE Internet of Things Journal*, 5(3), 1516–1526. doi:10.1109/JIOT.2017.2779820

Yu, C., Yu, L., Wu, Y., He, Y., & Lu, Q. (2017). Uplink scheduling and link adaptation for narrow-band Internet of Things systems. *IEEE Access: Practical Innovations, Open Solutions*, 5, 1724–1734. doi:10.1109/ACCESS.2017.2664418

KEY TERMS AND DEFINITIONS

Applications of NB-IoT: Uses of the NB-IoT technologies are diverse. Collectively it forms its applications. Several applications are available and new applications are emerging every year.

Energy Efficient IoT: Energy efficient IoTs reduce the consumption of energy to a large extent. These IoTs utilize the available energy optimally and can also harness it locally if appropriate arrangements are in place.

Green IoT: Energy efficient IoT is known as Green IoT. NB-IoT is an example of green IoT. These IoTs are in high demand because the battery life for these IoTs is more than 10 years.

Internet of Things: It is the extension of the internet through the interconnection of sensors/ sensing devices which can provide a lot of information about the environments where they are deployed. They can also activate or deactivate several operations through appropriate actuators.

Narrowband Internet of things (NB-IoT): NB-IoT is a specific type of IoT which uses a very small bandwidth and transmit power for its operations. It is a low power wide area network technology. It has been standardized in LTE Release 13 and 14.

NB-IoT Deployment: Placement of sensors, transceivers, actuators and other components of NB-IoT at the appropriate locations and the allocation of resources is collectively known as its deployment. Successful deployment is a challenging issue for any IoT.

Interoperability Provision of IoT Data Protocols on Top of Virtualized Infrastructure

Vaios Koumaras

INFOLYSiS P.C., Greece

Marianna Kapari

INFOLYSiS P.C., Greece

Angeliki Papaioannou

INFOLYSiS P.C., Greece

George Theodoropoulos

INFOLYSiS P.C., Greece

Ioannis Stergiou

INFOLYSiS P.C., Greece

Christos Sakkas

INFOLYSiS P.C., Greece

Harilaos Koumaras

NCSR “Demokritos”, Greece

INTRODUCTION

Internet of Things (IoT) focuses on the interconnectivity amongst people, sensors, actuators and processes. This term aims on the connection of various devices in order to provide data analytics and automation (Haseeb et al, 2017). An important requirement of IoT is ubiquitous connectivity and in order to achieve it, applications are needed for supporting the various sets of devices that will be used, as well as, communication protocols (Buyya & Vahid Dastjerdi, 2016). For instance, the aforementioned devices can be a smartwatch, a fitness band (Ranger, 2018), a television, a car and many more IoT related end user devices (Rouse, 2018). In addition, the IoT idea is also based on the concept of a digital home in which various daily devices (for example appliances) go online, are interconnected, and allow efficient data exchange among them.

Therefore, various sensing devices like Radio Frequency Identification Devices (RFID), scanners, GPS and so on are combined in order to form a wider network. The aim of that wider network is to have various devices connected and make possible the identification and management of them (Liu & Lu, 2012). Due to the variety of devices being connected together in order to allow data exchange, interoperability issues are encountered, which limit the applicability of IoT-related services due to the variety of IoT data protocols utilized by different sensors and data nodes. This article addresses the issues of IoT

DOI: 10.4018/978-1-7998-3479-3.ch064

interoperability and possible solutions in relation to it. Through the agility brought by virtualization, which allows gateway functions to be deployed on demand in an agile manner, this chapter discusses the concept of interoperability provision over heterogeneous IoT data protocols.

More specifically, this chapter discusses the problem of interoperability in IoT domains by introducing the most popular IoT data protocols that are commonly used today. Then, a proposed solution is described based on the agility offered by virtualized infrastructures with SDN/NFV capabilities. The basic functionalities of these technologies are briefly presented and finally a proposed approach towards the deployment on-demand of data protocol GWs is described.

BACKGROUND

During 1980s and 1990s, there was the concept of making objects smart, by integrating to them processing and connectivity capabilities, but it was slowed down due to the low progress in technological advancements (Ranger, 2018). Additionally, in the book “When Things Start to Think”, professor Neil Gershenfeld used the term “Things” which shows that there was an initial vision of Internet of Things (IoT) concept, back in 1999 (Rouse, 2018). The term IoT was coined by Kevin Ashton in a presentation regarding supply-chain management in 1999 (Kramp 2013). While he was part of the Auto-ID Center at MIT, he was involved in the Radio Frequency Identification Devices (RFID) applications extension in broader domains which is the basis of the Internet of Things (Buyya & Vahid Dastjerdi, 2016). Kevin Ashton with the term IoT wanted to describe a system that the real world is being connected using a ubiquitous network of sensors (Corcoran, 2016).

In order to fully understand the concept of IoT is better to consider the Internet, which is the basic infrastructure for interconnecting remote networking nodes, such as the IoT objects. The Internet origins are dated back to 1960s. TCP/IP was introduced in 1974 and by 1984 a thousand nodes were switched over to it for data transmission and as a networking protocol. In the 1990s and early 2000s, this connectivity trend continued, which gradually resulted in numerous connected terminals and devices. In present-day, people use smartphones in order to perform their daily tasks, mobile network allows for ubiquitous connectivity and the internet allows device connection from anywhere (Corcoran, 2016).

Nowadays, IoT applications have been found in different sectors, such as to systems which are industrial and closed-loop to commercial ones (Kolias, Stavrou, Voas, Bojanova & Kuhn, 2016). Examples of the devices are: cameras, lights, television, printers and so on. The number of devices which are currently connected to the internet is estimated to be around 5 billion and an increase is expected in 2020 to be around 25 billion. The future shows that IoT devices will be mainstream, increase technological advancements to areas such as healthcare by using monitoring wearables, to retail as well as transportation. There is a transition of the IoT devices and from monolithic boards are turning to modular devices so as to satisfy daily life's needs (Kolias, Stavrou, Voas, Bojanova & Kuhn, 2016).

IoT devices can use a variety of networks such as LAN, WAN, PAN or cellular networks. In addition, due to device restriction requirements such as battery, high autonomy and low cost, wireless technologies, such as RFID, Bluetooth, NFC are used (Díaz Zayas, García Pérez, Recio Pérez & Merino, 2017). RFID takes an important role in IoT. The RFID system includes back-end network applications and data acquisition. Devices as such even though they play an important role in IoT, still face various problems like standards compatibility. Moreover, besides RFID, wireless networks play an important role, allowing users to connect via various devices, creating the basis for the wider deployment of IoT systems (Liu & Lu, 2012).

IoT DATA PROTOCOLS

The devices, which are connected, have different protocols, which lead to interoperability issues (Blackstock & Lea, 2014) when they need to exchange data which are heterogeneous protocol-wise. Examples of protocols that are usually utilized by IoT sensors and devices as communication data protocols are the MQTT, CoAP, HTTP and UDP. These protocols are briefly described in the following sections.

MQ Telemetry Transport protocol (MQTT)

MQ Telemetry Transport (MQTT) is a lightweight messaging protocol, which is designed for constrained devices, networks which are not sufficiently reliable, namely utilizing low bandwidth with high latency (MQTT, 2014). According to Rouse (2018), it is a lightweight messaging protocol, since its messages consist of a fixed header which is 2 bytes, an optional variable header, a QoS (Quality of Service) level and a theoretically max message payload to 256Mb, but the max payload size supported by IoTF is 4Kb. MQTT's basic principles are to minimize the network bandwidth as well as the requirements of the device and ensure reliability and assurance of delivery. It is also ideal for machine to machine or IoT situations where various devices are connected as well as in cases where mobile applications are used and bandwidth and battery power are important. The term MQTT was coined in 1999 by Dr Andy Stanford-Clark from IBM and Arlen Nipper of Arcom (MQTT, 2014). MQTT has a diverse structure which can be a huge tree with no clear way to be able to divide it in smaller components that can be manageable. Therefore, this is a problem which in turn increases the complexity of the network. In addition, MQTT protocol has a problem with interoperability. A problem in terms of interoperability is due to the binary message payloads which contain no information of their encoding. Problems as such arise when applications from different manufacturers need to be connected and work together. Moreover, another problem is that the receiver is not able to know who the original message sender is. The only way to be able to know the sender is if it is mentioned in the message. The use of additional security features on top of the MQTT protocol causes a code increase which leads to difficulty handling the implementations (Rouse, 2018).

Constrained Application Protocol (CoAP)

Constrained Application Protocol (CoAP) is a web transfer protocol which is used with constrained nodes and networks. It is designed to be used with machine to machine applications like smart energy or building automation (Bormann, 2016). In specific, it is a specialized web transfer protocol for use with constrained nodes and constrained networks in the Internet of Things. The protocol is designed for machine-to-machine (M2M) applications such as smart energy and building automation.

In addition, CoAP allows devices like low power sensors and actuators to be able to communicate via the Internet. This protocol runs on devices which support UDP (or else User Datagram Protocol) and is ideally suited for devices which are of low memory and power as it offers small message size, lightweight messages and message management (Johnson, 2016).

This protocol, like the HTTP protocol, is based on the REST model which allows servers to make resources available under a URL. The clients can access those resources by using methods/commands like GET, POST, PUT and DELETE (Bormann, 2016). A description of the aforementioned methods can be found below (Shelby, Hartke & Bormann, 2014).

- GET: This method allows the retrieval of a representation in relation to information of the resources identified by the URI. Get is a safe method to be used.
- POST: This method refers to the processing of the enclosed representation. POST is neither a safe method nor an idempotent one.
- PUT: This method refers to the update or creation of resources identified by URI. PUT is neither a safe nor an idempotent method like POST.
- DELETE: This method allows the deletion of information identified by the URI. It is not safe a safe method but is idempotent.

The CoRE group has designed CoAP with the following features in mind:

- Overhead and parsing complexity.
- URI and content-type support.
- Support for the discovery of resources provided by known CoAP services.
- Simple subscription for a resource, and resulting push notifications.
- Simple caching based on max-age.

Hypertext Transfer Protocol (HTTP)

Hypertext Transfer Protocol (HTTP) is a network protocol that web browsers and servers use in order to communicate. The HTTP can be easily recognized in a web browser by looking at the URL as it has the signature of http in the beginning. This protocol is similar to other protocols like FTP (File Transfer Protocol) as it uses a program in order to request files from a server remotely. In the case of HTTP, the program that is used to request files is a web browser and requests HTML files which are displayed in the format of text, images, hyperlinks and so on. In addition, this protocol unlike other protocols is stateless, which refers to the ability that it has to stop the connection once a request has been made and executed. This protocol has been created by Tim Berners- Lee in the early 1990s (Mitchell, 2018).

HTTP offers three message types which are GET, POST and HEAD. The GET type refers to messages which are sent to a server and contain only a URL. The URL can have zero or more data parameters which can be added at the end of it (Mitchell, 2018).

The POST type refers to any additional data parameter which is placed in the body of the message and not in the end of the URL (as it happens in the case of the GET type). The HEAD type is similar to GET but instead of replying with full URL information, the server sends back only information in regards to the header (Mitchell, 2018).

As with the other protocols, the HTTP has its issues. Some of the issues in relation to this protocol relate to message transmission failure due to different reasons like:

- User error
- A problem with the web server or browser
- An error that exists in the web pages creation and
- Temporary problems in regards to the network

In case of an issue, the protocol captures the problem of failure and then reports an error code in order to point out what kind of error it is (Mitchell, 2018).

User Datagram Protocol (UDP)

User Datagram Protocol (UDP) is a transport layer protocol and is used with the IP network layer protocol. UDP is written by John Postel and defined by the RFC 768 and provides a datagram service to an end system (Fairhurst, 2008). The issue with this protocol is that it provides an unreliable service meaning that there is no guarantee that the message will be delivered and protected from duplication. It is an unreliable, minimal, message passing transport to applications as well as upper layer protocols. It offers an efficient communication transport but it does not have reliability or congestion control. Due to its unreliability, application designers often consider it as a transport. In cases of data duplication, the designers need to verify that the application can handle duplication and implement appropriate mechanisms in order to detect it. Moreover, this protocol has security issues. In cases of applications that have sensitive data, a separate security service needs to be provided in order to use additional protocol mechanisms (Fairhurst, 2008).

IoT ARCHITECTURE COUPLED WITH CLOUD COMPUTING SYSTEM

The architecture of a modern IoT system today is usually coupled with cloud computing systems, including both central and edge nodes, providing increased scalability and performance, as well as third party access to the IoT infrastructure. The first layer consists of sensors and actuators. The second layer (Internet Gateway) involves sensor data aggregation systems and a conversion of analog to digital data. The third layer (Edge IT and Fog computing systems) consists of IT systems which preprocess data prior to its transfer to a data center or a cloud. The last layer (Data Center/Cloud), layer four, is the stage where data can be analyzed, managed and stored. Below follows a description of each of these IoT stages.

Figure 1. Architectural Layers of a modern IoT system

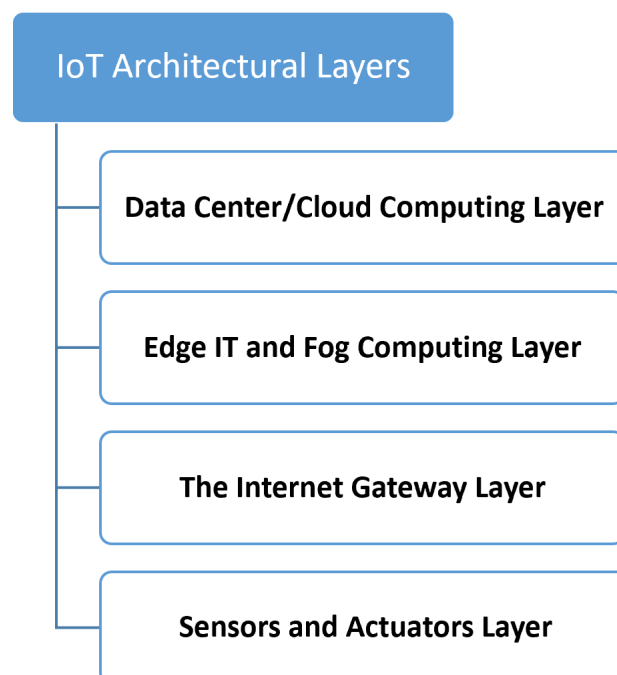


Figure 1 depicts the four layers of an IoT architecture couple with cloud computing resources, which are briefly described hereby:

- **Sensors and Actuators:** The sensors have the ability to collect data from the environment and provide useful data. A different name of a sensor is transducer which is a physical device that is used so as to convert one energy form to another one. An example of a sensor is a microphone that can take the sound waves and convert them into electrical energy. (Eller, 2017)

A transducer that can be found in different IoT systems is an actuator, which means that it operates in an opposite direction of a sensor. An actuator takes the electrical input and converts it to a physical (Eller, 2017). Examples of an actuator's process are: to shut off the power supply, to adjust the air flow valve and so on. The first stage of sensors/actuators involves different devices such as accelerometers, air sensors, heart rate monitors and so on (Fuller, 2018)

- **The Internet Gateway Layer:** Data from sensors need to be aggregated to one node for further processing. Ccssioanlly and depending on the node type, the data may need to change before it enters the processing stage. Moreover, data need to change from its analog form to a digital one and the Internet Gateway Layer assists in achieving this. Sensors/Actuators and the Internet Gateway Layer can be found close to each other and therefore, that is helpful in order to achieve real time pre-processing. Moreover, the internet gateways can be built and support analytics, malware protection and data management (Aher, 2018). A gateway is an edge device, it is located externally to the data center and it is important to know its location. Additionally, gateway devices can be found in various environments like factory floor to mobile field stations and they are built in order to be portable, durable to temperature changes, vibration and deployable (Fuller, 2018).
- **Edge IT and Fog Computing:** This is the third stage of the IoT architecture. In this stage, data enter in the IT systems so as to be further analyzed. This layer (Edge IT) is located in remote offices or in edge locations comparing to the two aforementioned layers (Sensors/Actuators and the Internet Gateway Layer) which are located close to each other. The IoT data can be huge in size thus taking a big amount of network bandwidth and due to this reason, the edge systems via analytics allow data to be minimized in terms of its amount (Aher, 2018). In addition, while processing the data different issues can be encountered such as security issues, storage issues and data delay. By following a staged approach, the data can be preprocessed, meaningful results can be generated and only those will be passed on (Fuller, 2018).
- **Data Center/Cloud:** This is the fourth and last stage of the IoT architecture. Data can be forwarded to cloud-based systems or a physical data center when they need to be further processed. In this way, data will be analyzed, managed and stored by powerful IT systems (Fuller, 2018). Moreover, by exploiting the virtualization capabilities that cloud systems offer, additional functions may be also considered in the processing process in order to perform additional tasks, such as big data analytics or business intelligence tasks.

The generic architecture of IoT systems consists of systems which are complex and perform various tasks such as collecting information in relation to people, environment, psychological measurements, tracking things and so on. Information as such is stored in IoT back end with or without the aid of a gateway (Haseeb, Hashim, Khalifa, Ismail, 2017). A gateway stores, translates and forwards data to an IoT back-end system. There are also gateways which are more complex and allow the proprietary device

management of IoT devices. In addition, there are IoT gateways which allow process automation prior to forwarding the processed data to a back- end cloud for storage and additional processing. The IoT back-end has a limit of the capabilities of a physical IoT device and does not have a mechanism in order to overcome this issue. (Haseeb, Hashim, Khalifa, Ismail, 2017). The coupling of traditional IoT systems with cloud computing and virtualization infrastructure can contribute to overcome the aforementioned limitations by adding functions in an agile manner that are able to perform tasks on demand. Among the various functions, one of the main challenges that the modern gateways have to face is the interoperability provision over the various data protocols that the IoT sensors and nodes use.

IoT INTEROPERABILITY: VERTICAL AND HORIZONTAL DEPLOYMENTS

An important requirement of IoT is interoperability as the “things” which are connected to the Internet need also to be found, accessed and managed in a unified way. Therefore, interoperability plays an important role in this case and is a feature that it should be appropriately supported by the IoT Gateway.

In an IoT architecture, the devices need to interoperate and communicate with a gateway which is cloud based. This though is not always true due to the various devices which are connected, many communication protocols which are used as well as many application layer interfaces. Thus, this diversity of devices leads to vertical solutions which are isolated. These vertical solutions don't allow systems to communicate in order to exchange valuable information and therefore do not follow the true nature of IoT (Haseeb, Hashim, Khalifa, Ismail, 2017).

Figure 2. Vertical IoT approach

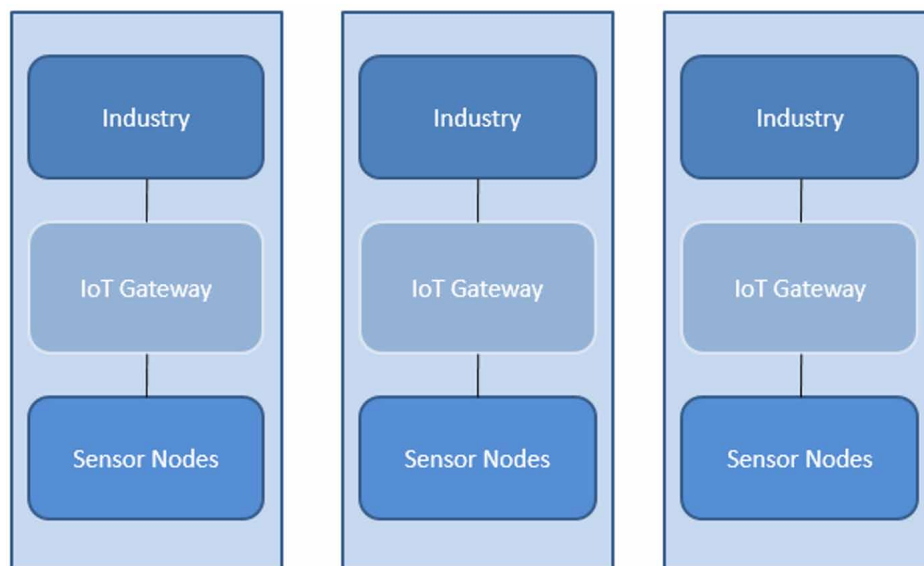
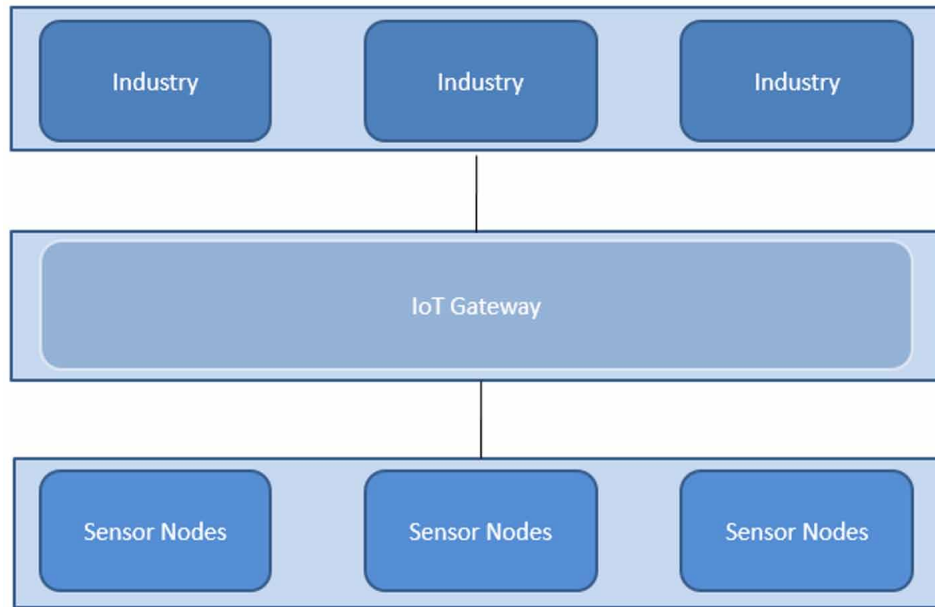


Figure 2 depicts a graphic of the vertical isolated approach. Vertical solutions have dominated the IoT devices and systems. A company using a vertical model can control an IoT device, a gateway, an application and also a cloudbased service. Additionally, there is a clear control in terms of decision making from the entry stages and also the users deal with one provider in cases of technical difficulties.

The vertical solutions have more drawbacks than benefits, specifically in terms of user perspective. An example is that there may be different companies offering different solutions to a building (one company can focus on remote control on electric light usage and the other on the security system's IoT sensor) thus this is impossible to control the devices using a single application as each of the companies use their own platform (IoT, 2018). In addition, it is in a virtual manner impossible to be able to manage a diverse set of IoT devices (Haseeb, Hashim, Khalifa, Ismail, 2017).

Figure 3. Horizontal IoT approach



Besides the vertical solution, there is the horizontal approach. Figure 3 shows an example of an horizontal solution. A horizontal platform is used so as to streamline the user control of IoT devices. In the horizontal approach, there is a platform which allows various companies to be able to develop applications, IoT devices, which are compatible and also to interact amongst each other using a common framework. Furthermore, this approach is an ideal choice for enterprises while it offers ease of use, application development. The horizontal solution reduces system costs and increases the data monetization opportunities (IoT, 2018).

IoT INTEROPERABILITY PROVISION BASED ON VIRTUALIZED INFRASTRUCTURE

IoT offers interconnectivity amongst diverse devices but this comes with a price as there is a lack of interoperability. Due to the reason that the companies offer solutions independently of each other, have different platforms and frameworks, there is a problem with integration with one another (Eastwood, 2017). In order to overcome the interoperability issue, a solution is by using virtualization and software-defined techniques. There are two technologies which promise a cost effective scale and versatility,

which is necessary for IoT, namely the Software Defined Networking (SDN) and Network Functions Virtualization (NFV) (Bizanis, Kuipers, 2016).

Software Defined Network

SDN started on campus while researchers were experimenting with new protocols. The result was to come up with an idea of making network devices programmable and allowing them to be controlled by a central element. In addition, SDN allows: i) The separation of control and forwarding function, ii) Control centralization, and iii) Has the ability to program the network's behavior with the use of well-defined interfaces (Pate, 2013).

In SDN, the IoT elements are devices which have no intelligence in order to control as well as forward data traffic. This technology acts as the brain of an entire network and resides on multiple servers in a cloud network. Additionally, the entire network management and the control of operations are called the SDN controller. SDN controls the network and reduces the burden of network operators as it aims on avoiding network configuration errors which is an issue very common today (Alenezi, Almustafa, Meerja, 2018).

Network Functions Virtualization

Network Functions Virtualization (NFV) was created by a service providers' consortium. The service providers tried to speed up the network services so as to increase revenue and growth plans and found that hardware-based appliances had a limit in order to achieve that. So, NFV was a solution in order to accelerate service innovation and provisioning (Pate, 2013). In NFV, there is a software manager similar to SDN in order to allocate and release resources. In addition, NFV devices can be controlled by using a central SDN controller thus making it easier for SDN and NFV to coexist and function together. The NFV architecture is composed of three components: Physical Hardware, Virtual Hypervisor Layer and Virtual Machine (Alenezi, Almustafa, Meerja, 2018). The Virtual Hypervisor Layer runs on hardware and manages resources like the power of CPU, memory and storage capacity. The Virtual Machine refers to software, which has the ability to emulate the architecture and functionalities of the physical platform by using hardware resources. A physical hardware is able to host more than one virtual machines. (Alenezi, Almustafa, Meerja, 2018). NFV has been a popular choice amongst telecommunication companies or else Telcos as they are suffering from interoperability issues. NFV aimed at aiding on operational challenges, high costs, device heterogeneity and proprietary appliances. Issues regarding heterogeneous devices are the following (Haseeb, Hashim, Khalifa, Ismail, 2017):

- 1) Fixed configuration: The Telco hardware has a fixed IP which does not change for a long period of time
- 2) Manual Management: The management and configuration of telco equipment requires moving physical staff and this can be achieved once at a time.
- 3) Rapid growth of IP end points: Many users are able to access the Internet thus this makes it difficult if a central provisioning is not established.
- 4) Network endpoint mobility: Due to the fixed nature of networks, the requirements for mobility take a longer time so as to reconfigure the network in regards to different scenarios.
- 5) Elasticity: Due to the reason that there is no mechanism in order to upgrade or even downgrade physical software, the networks need to be over provisioned.

Software Defined Network and Network Functions Virtualization

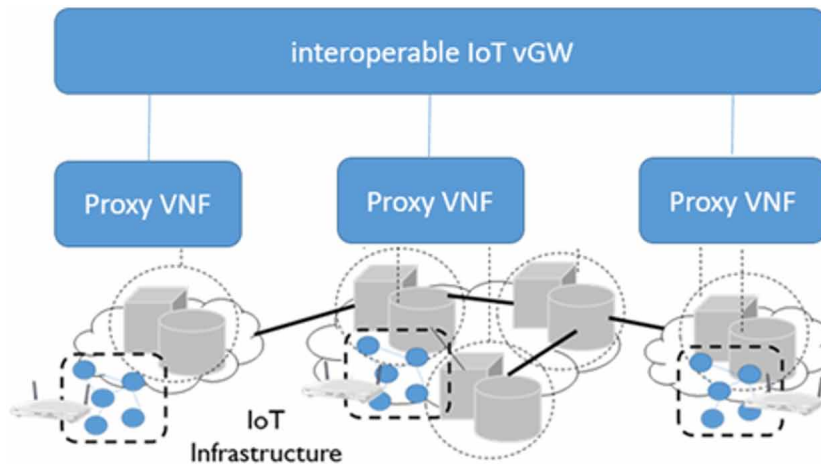
These two technologies are quite different but they share common principles. The WAN or else Wide Area Network is an environment which is closed and static and even though improvements in regards to automatic capabilities have been made, there is still a manual process in terms of general management and delivery. The use of SDN and NFV allows the networks to be driven by centralized software control capabilities which are able to program an entire network using a set of open APIs. Virtualization is also beneficial as the costs of deploying services around the network minimize (Vodafone, 2016). NFV is driven by dynamic requirements and is associated with the evolution of cloud. Nowadays, there is an evolution in terms of delivery service thus virtual technologies offer the same services as physical ones and they are also instantiated when they are needed. NFV offers a dynamic set of capabilities and requires a dynamic network in order to connect them. SDN takes many forms like in the example of SD-WAN which targets an enterprise WAN. A challenge of SDN is that the network is controlled using open APIs without affecting the physical underlying connectivity. A customer will be able to see development in terms of self service capabilities which are not easily available in the current day in terms of WAN and security services. In addition, another positive aspect is to provide real-time visibility in terms of transport and application performance across the network. In order to decide whether to deploy SDN or not, an organization needs to understand several aspects such as the sites connected to the network, the application flows and the data type that these sites will traverse. When the previous aspects have been decided and understood, then it will be easier for the business to consider the dynamicity of its services, the criticality of data and the internet transport that the content needs to consume (Vodafone, 2016).

Furthermore, the use of these technologies (SDN and NFV) aims to face the 5G interoperability challenge by allowing the automatic deployment and programming of network services. These technologies are cost effective and provide the necessary versatility for the IoT services. Additionally, the 5G networks need to support software network technologies and the SDN and NFV are important as they are an ideal choice. As mentioned previously the SDN aims at basic connectivity services whereas the NFV provides orchestration, a virtualization framework and aims at a larger scope. In order to be able to implement these technologies a virtual gateway is needed as it ensures the interoperability amongst objects and technologies. Furthermore, additional KPIs metrics need to be evaluated in order to ensure the quality of QoS and QoE which is perceived by the end user. The IoT platform needs to be consisted of physical and virtual parts and be on top of a testbed so that the protocols MQTT, CoAP UDP will be instantiated and set up in order to provide an interoperable layer by using a virtual gateway.

The virtual gateway's aim is to provide the interoperability of IoT NFVs in different domains so as to unify the various and diverse smart objects which operate by using different IoT protocols. The ideal choice of protocols is MQTT and CoAP as they are open standards and they provide a more suitable solution to constrained environments in regards to HTTP.

In specific (see Fig. 4), the project envisages the testing/deployment of specific VNFs, which are mapping functions (i.e. proxies) of popular IoT data protocols (such as MQTT and CoAP) to generic data protocol (such as UDP or HTTP) in order to provide a common interoperable layer over the heterogeneous IoT nodes deployed at heterogeneous IoT domains and testbeds.

Figure 4. SDN/NFV-driven IoT interoperability on top of a virtual-based GW



FUTURE RESEARCH DIRECTIONS

The VNF formulation of the software-based IoT mapping functions can facilitate further the SaaS business model of them, while the integration of SDN rules in the virtualization environment can further upgrade the provision of the interoperable services following the SaaS/PaaS model. A major factor towards this in the future is the provision and support of slicing, which will support that service isolation can be achieved and therefore multi-tenancy exploitation of the infrastructure.

CONCLUSION

Internet of Things aims at the interconnectivity between people, sensors, actuators and processes. Moreover, diverse devices from various vendors are being used and connected in order to provide data analytics and automation. IoT allows everyday objects to act like “smart” and being connected to the internet and allow the users to perform various tasks. Examples of IoT can be a smartwatch, a fridge, a car and so on. Due to the vast majority and diversity of devices which are being connected and share information, as well as the different protocols each of them uses, a problem of interoperability is being defined. This interoperability issue can be solved by using virtualization technologies such as SDN and NFV. These virtualization techniques allow the minimization of costs that existed in physical devices, share same services as the physical ones but they are instantiated when they are needed. In order to use these technologies at their best possible way, it is important to consider several aspects before proceeding on choosing them such as the sites that will be connected to the network, the application flows and the data type that these sites will traverse. Furthermore, the combination of SDN and NFV aims to face the interoperability challenge as these technologies are cost effective and provide versatility for the IoT services. In order to implement these technologies, a virtual gateway is needed so as to achieve the interoperability between the objects and technologies. The IoT platform consists of physical and virtual parts and is on top of a testbed so that the protocols MQTT, CoAP and UDP will be instantiated.

REFERENCES

- Aher, B. (2018). *A Look at the IoT Architecture*. Retrieved 8 24, 2018, from DZone: <https://dzone.com/articles/iot-architecture-2>
- Alenezi, M., Almustafa, K., & Meerja, K. A. (2018). *Cloud based SDN and NFV architectures for IoT infrastructure*. Egyptian Informatics Journal., doi:10.1016/j.eij.2018.03.004
- Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., ... Warfield, A. (2003). Xen and the Art of Virtualization. *Proc. 19th ACM Symposium on OS Principles*.
- Bavier, A., Feamster, N., Huang, M., Peterson, L., & Rexford, J. (2006). In VINI Veritas: Realistic and Controlled Network Experimentation. *Proc. ACM SIGCOMM*.
- Bhatia, S., Motiwala, M., Muhlbaier, W., Mundada, Y., Valancius, V., Bavier, A., ... Rexford, J. (2008). Trellis: A Platform for Building Flexible, Fast Virtual Networks on Commodity Hardware. *Proc. 3rd ACM Workshop on Real Overlays and Distributed Systems*. 10.1145/1544012.1544084
- Bizanis, N., & Kuipers, F. A. (2016). SDN and Virtualization Solutions for the Internet of Things: A Survey. *IEEE Access: Practical Innovations, Open Solutions*, 4, 5591–5606. doi:10.1109/ACCESS.2016.2607786
- Blackstock, M., & Lea, R. (2014). IoT interoperability: A hub-based approach. *2014 International Conference On The Internet Of Things (IOT)*. doi: 10.1109/iot.2014.7030119
- Bormann, C. (2016). *CoAP — Constrained Application Protocol | Overview*. Retrieved from <http://coap.technology>
- Boutaba, N., & Chowdhury, R. (2009). Network virtualization: State of the art and research challenges. *IEEE Communications Magazine*, 47(7), 20–26. doi:10.1109/MCOM.2009.5183468
- Buyya, R., & Vahid Dastjerdi, A. (2016). *Internet of things*. Cambridge, MA: Morgan Kaufmann is an imprint of Elsevier.
- Carapinha, J., & Jimenez, J. (2009). Network Virtualization – a View from the Bottom. *Proc. ACM SIGCOMM VISA'2009 Workshop*.
- Chowdhury, M., Rahman, M., & Boutaba, R. (2012). ViNEYard: Virtual Network Embedding Algorithms With Coordinated Node and Link Mapping. *IEEE/ACM Transactions on Networking*, 20(1), 203–226. doi:10.1109/TNET.2011.2159308
- Corcoran, P. (2016). The Internet of Things: Why now, and what's next? *IEEE Consumer Electronics Magazine*, 5(1), 63–68. doi:10.1109/MCE.2015.2484659
- DíazZayas, A., García Pérez, C., Recio Pérez, Á., & Merino, P. (2017). 3GPP Evolution on LTE Connectivity for IoT. *Internet of Things*, 1-20. doi:10.1007/978-3-319-61300-0_1
- Eastwood, G. (2017). *IoT's interoperability challenge*. Retrieved from <https://www.networkworld.com/article/3205207/internet-of-things/iots-interoperability-challenge.html>
- Eller, J. (2017). *IoT System | Sensors and Actuators*. Retrieved August 24, 2019, from Bridgera: <https://bridgera.com/iot-system-sensors-actuators/>

Fairhurst, G. (2008). *The User Datagram Protocol (UDP)*. Retrieved from <https://www.erg.abdn.ac.uk/users/gorry/course/inet-pages/udp.html>

Fremantle, P. (2015). *A Reference Architecture For The Internet of Things*. Retrieved from WSO2: <https://wso2.com/whitepapers/a-reference-architecture-for-the-internet-of-things/#04>

Fuller, J. (2018). *The 4 stages of an IoT architecture*. Retrieved from TechBeacon: <https://techbeacon.com/4-stages-iot-architecture>

Haseeb, S., Hashim, A., Khalifa, O., & Ismail, A. (2017). *Connectivity, interoperability and manageability challenges in internet of things*. doi:10.1063/1.5002022

Infolyxis. (2018). *Internet of Things*. Author.

IoT. (2018). *Defining the IoT Ecosystem for Enterprises*. Retrieved from IoT Innovation: <https://internet-of-things-innovation.com/insights/the-blog/defining-iot-ecosystem-enterprises/>

Johnson, S. (2016). *Constrained Application Protocol: CoAP is IoT's 'modern' protocol*. Retrieved from: <https://internetofthingsagenda.techtarget.com/feature/Constrained-Appliation-Protocol-CoAP-is-IoTs-modern-protocol>

Kolias, C., Stavrou, A., Voas, J., Bojanova, I., & Kuhn, R. (2016). Learning Internet-of-Things Security “Hands-On”. *IEEE Security and Privacy*, 14(1), 37–46. doi:10.1109/MSP.2016.4

Kramp, T., van Kranenburg, R., & Lange, S. (2013). Introduction to the Internet of Things. In *Enabling Things to Talk*. Springer. doi:10.1007/978-3-642-40403-0_1

Liu, T., & Lu, D. (2012). The application and development of IOT. *2012 International Symposium On Information Technologies In Medicine And Education*. doi: 10.1109/itime.2012.6291468

Masti, S., & Raghavan, S. (2012). VNA: An Enhanced Algorithm for Virtual Network Embedding. *21st International Conference on Computer Communication Networks (ICCCN)*. 10.1109/ICCCN.2012.6289180

McKeown, T. A. (2008). *Openflow: enabling innovation in campus networks*. SIGCOMM CCR. doi:10.1145/1355734.1355746

Mitchell, B. (2018). *What Does HTTP Mean?* Retrieved from <https://www.lifewire.com/hypertext-transfer-protocol-817944>

MQTT. (2014). Retrieved from Oasis: <http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html>

Network Function Virtualization (NFV) based architecture to address connectivity, interoperability and manageability challenges in Internet of Things (IoT). (2019, August 11). *IOP Conference Series. Materials Science and Engineering*, 260, 012033. doi:10.1088/1757-899x/260/1/012033[Accessed

Nogueira, J., Melo, M., Carapinha, J., & Sargento, S. (2011). A platform for Operator-driven Network Virtualization. *Proc. IEEE EUROCON - International Conference on Computer as a Tool*. 10.1109/EUROCON.2011.5929325

Nogueira, J., Melo, M., Carapinha, J., & Sargento, S. (2011). Network Virtualization System Suite: Experimental Network Virtualization Platform. *International Conf. on Testbeds and Research Infrastructures for the development of Networks and Communities*.

Nogueira, J., Melo, M., Carapinha, J., & Sargento, S. (2011). Virtual Network Mapping into Heterogeneous Substrate Networks. *Proc. IEEE ISCC -11*. 10.1109/ISCC.2011.5983876

Pate, P. (2013, March 30). *NFV and SDN: What's the Difference?* Retrieved from <https://www.sdxcentral.com/articles/contributed/nfv-and-sdn-whats-the-difference/2013/03/>

Ranger, S. (2018). *What is the IoT? Everything you need to know about the Internet of Things right now* | *ZDNet*. Retrieved from <https://www.zdnet.com/article/what-is-the-internet-of-things-everything-you-need-to-know-about-the-iot-right-now/>

Rouse, M. (2018). *What is internet of things (IoT)? - Definition from WhatIs.com*. Retrieved from <https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT>

Rouse, M. (2018). *What is MQTT (MQ Telemetry Transport)? - Definition from WhatIs.com*. Retrieved from <https://internetofthingsagenda.techtarget.com/definition/MQTT-MQ-Telemetry-Transport>

Shelby, Z., Hartke, K., & Bormann, C. (2014). *RFC 7252 - The Constrained Application Protocol (CoAP)*. Retrieved from <https://tools.ietf.org/html/rfc7252>

Vodafone. (n.d.). *Software Defined Networking and Network Function Visualization*. Retrieved from www.vodafone.com/business/news-and-insights/blog/gigabit-thinking/software-defined-networking-and-network-function-virtualisation

Werle, C., Papadimitriou, P., Houidi, I., Louati, W., Zeghlache, D., Bless, R., & Mathy, L. (2011). Building Virtual Networks Across Multiple Domains. *Proc. ACM SIGCOMM 2011*. 10.1145/2018436.2018495

Yang, M., Y. L. (2013). OpenRAN: A Software-defined RAN Architecture Via Virtualization. *Proc. SIGCOMM*. 10.1145/2486001.2491732

Yu, J. (2012). Solution for virtual network embedding problem based on Simulated Annealing genetic algorithm. *2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*.

Zhang, S., Qian, Z., Guo, S., & Lu, S. (2011). FELL: A Flexible Virtual Network Embedding Algorithm with Guaranteed Load Balancing. *Proc. 2011 IEEE International Conference on Communications (ICC)*. 10.1109/icc.2011.5962960

Zu, Y., Zhang-Shen, R., Rangarajan, S., & Rexford, J. (2008). Cabernet: Connectivity Architecture for Better Network Services. *Proc. ACM ReArch '08*.

KEY TERMS AND DEFINITIONS

IoT: Internet of things. The interconnection via the internet of computing devices embedded in everyday objects, enabling them to send and receive data.

IoT Interoperability: The ability for IoT systems or components of systems to communicate with each other, regardless of their manufacturer, technical specifications, or data protocols.

IoT Protocols: A standard set of rules that allow IoT devices to communicate with each other. These rules include what type of data may be transmitted, what commands are used to send and receive data, and how IoT data transfers are confirmed.

Network Virtualization: Abstracting network resources traditionally delivered in hardware to software. It can combine multiple physical networks to one virtual, software-based network, or it can divide one physical network into separate, independent virtual networks.

NFV: Network function virtualization. Network architecture concept that uses the technologies of IT virtualization to virtualize entire classes of network node functions into building blocks that may connect, or chain together, for creating/providing communication services

SDN: Software-defined networking. SDN architecture aims to make networks agile and flexible. Dynamic, programmatically efficient network configuration in order to improve network performance and monitoring making it more like cloud computing than traditional network management.

Network Science for Communication Engineering

7

Sudhir K. Routray

 <https://orcid.org/0000-0002-2240-9945>

Bule Hora University, Bule Hora, Ethiopia

INTRODUCTION

Network science is associated with the study, analysis, modeling and applications of networks. Networks are not limited to any one or two disciplines. Rather, networks are found in almost every field in different forms. For instance, in electrical engineering we have electrical circuits which are known as networks. In communication engineering, we have communication networks of different types such as access networks, regional networks, core networks and the Internet. Computer networks are very much identical to the communication networks these days. In fact, unification of communication and computing has given birth to the Internet. Power distribution networks are found throughout the world. They are huge in size and directly affect the human activities across the world these days. In our society, we have different types of social networks. Social networks through the communication platforms have given birth to the online social networks such as LinkedIn and Facebook. However, the social networks are very much older than the current social networks found online. The social networks were started along with the civilizations. However, their studies in psychology were started much later in the twentieth century (Moreno, 1934). Transportation networks are formed through the roads connecting cities, towns and villages. These transportation networks are also found over the water and air as the shipping and airline networks. Based on the economic activities between different countries and regions, trade and economic networks are formed.

In the nature, many different types of networks are found in various forms. Human brain is a complex network of neurons. In the human body there are some other networks as well such as blood circulation network and the lymphatic networks. In genetics, we can find out the link between generations and different clans. It is also a complex network. Based on the genetics, doctors classify different diseases and utilize them for advanced research and clinical treatments. This branch of medicine is commonly known as network medicine. Metabolism of organisms follows different sequences of processes which forms a metabolic network. Rivers and streams form natural water supply networks. In the modern world, there are several resource supply and distribution networks. Structural networks are found in different mechanical designs and civil architectures. In scientific collaborations we find the citation networks.

Complex systems are the large and randomly related with their components. The components of these systems are complexly related with each other. In other words, these systems are made up of many heterogeneous and non-identical elements connected through several diverse interactions. In every complexity, we always find a network is associated (Albert & Barabási, 2002). In order to understand the complexities, the underlying networks have to be understood properly. As we have seen above, every discipline has complex issues. Mapping the complexities accurately we can find a network which has the potential to simplify its understanding (Albert & Barabási, 2002). This is the main reason behind the popularity of network science.

DOI: 10.4018/978-1-7998-3479-3.ch065

Network Science: Origin and Background

Network science is being treated as the study of networks using different scientific methods. Its popularity and widespread applications are very much recent. However, the discipline has been developed quite clandestinely since a long time. The exact beginning of network related study is not known. Regular networks were being studied in geometry in several old civilizations. These principles were also being used in the building and structural architectures. However, graph theory which is considered as the mathematical basis of networks was started by Euler in 1735. His visit to Königsberg was the key in the development of graph theory (Alexanderson, 2006). The famous “Seven Bridges of Königsberg” puzzle was solved by Euler and it was the foundation of graph theory. Since then, graph theory became a systematic branch of mathematics. The first application of graph theory in engineering was found in the study and analysis of electrical networks in circuit theory. Early applications of network science are recorded in Psychology (Travers & Milgram, 1967) in which the psychologists tried to explain the social behavior of people. In the 1930s Moreno used the principles of networks to analyze the social networks (Moreno, 1934). The mathematical models for the networks other than the graph theory came quite late in the late 1950s. The mathematical model for the random networks was developed by Erdős and Renyi in 1959 (Erdős & Rényi, 1960). This was the first model to analyze the real networks. Their formulation assumed that every real network is random in nature as their nodal degrees follow Poisson’s distribution. In the late 1960s and early 1970s, communication networks were studied rigorously as their size and applications got expanded. The Granovetter network model was developed in 1973 which was very much fundamental like the random network model developed by Erdős and Renyi (Granovetter, 1977). Ecological networks and several biological networks were brought in to the picture in the late 1970s. After these practical applications, network science attracted researchers from almost every field. Two papers by the network scientists in the late 1990s opened the floodgate of research in this discipline. They are: the paper of Albert and Barabási (Albert & Barabási, 1999) in *Science*, on the scaling of the Internet published in 1999; and the paper of Watts and Strogatz (Watts & Strogatz, 1998) in *Nature*, regarding the small world networks published in 1998. These two papers showed the essence of network science in the modern scientific and technological analysis.

Since the publications of those two ground breaking papers in the late 1990s, network science became considered as a fundamental discipline in the study of networks and their applications to understand the natural complexities. Complex systems such as the Internet and human brain are built over large networks. In order to understand their principles, it is essential to understand their underlying network. Several branches of science, arts, and commerce now use network science as one of their main tools to analyze the complexities. For instance, in physics, chemistry, biology, communication engineering, material science, psychology, medicine, economics, and political science network science is used as a tool to explain the underlying complexities. In communication engineering, network science is used to study the dynamics and stability of the networks (Routray et al., 2015). The use of statistical modeling in the analysis of optical networks is found in Waxman (1988) in which the nodal degrees of the core networks were found to follow Poisson’s distribution. A complete statistical formalism for the fast estimation of optical network parameters was developed in Korotky (2004). A similar fast estimation of mesh optical networks was developed in Labourdette et al. (2005). Optical network topologies can be developed using the statistical principles of the optical networks (Pavan et al., 2010). Following the successes of these models several statistical models for different parameters of core and access networks have been developed in the recent years (Routray et al., 2015). Network economics can also be analyzed using network science approaches (Routray et al., 2017). The social networks are really complex and

their characteristics can only be understood when they are analyzed using mathematical models developed in network science.

Characteristics of Network Science

Now, network science has been established as a new branch of science which uses several scientific methods to analyze complex networks and complexities. It is an interdisciplinary field of study of networks to address the problems associated with large networks (Albert & Barabási, 2002). This field is very much empirical and data driven. It is founded on the logical bases of several mathematical concepts such as graph theory, probability, statistics and linear algebra (Goldenberg et al., 2010). It is also quantitative in nature because its parameters can be quantified. However, the estimations are very complex if the size of the network is large. Therefore, it needs computational tools for fast and accurate analysis of its parameters. Overall, it is a logical discipline which has the potential to resolve the complicated problems across many fields.

The main principles and applications of network science as observed in the recent times are: the applications of mathematics, in particular statistics and probability, to analyze the underlying concepts; to analyze the dynamics of the networks with the changing conditions; and to generate and design the networks found in the nature. The study of network science was started as the analysis of social interactions in the society (Moreno, 1934). In the 1950s, mathematicians developed analytical models to study the exact behaviors of the networks (Erdős & Rényi, 1960). In this work, the networks were considered as random and the nodal degree distributions normally follow Poisson's distribution. However, later it was found that this assumption of random behaviors of all real network nodal degrees was wrong. In Albert and Barabasi (1999), and Watts and Strogatz (1998), the new concepts of real complex networks came into light. Since then, network science has been systematized in a very scientific way. Some complex networks have been developed in the recent years such as the online social networks, huge trade networks based on the logistics, and the communication networks which is a complex hybrid of many different types of components. Newman et al. (2002) provided the basic framework for the analysis of social networks. Nekovee et al. (2007) and Borgatti et al. (2009) have developed the concepts for social network analysis and the anomalies in the behavior can point out the critical properties of the social networks. Applications of network science have spread in to several new areas which were very much uncharted before two decades. Network medicine is one such area in which the principles of network science are applied to medicine to find out complex relationships between the diseases (Barabási et al., 2011). Many such applications are emerging in almost all areas of science, arts, and commerce.

Network Science in Engineering

There are several complexities in engineering. However, there is no suitable tool to understand and analyze those complexities. In order to understand the complexities and to solve the problems linked with them, engineers have to first analyze and comprehend the associated networks. Many engineering systems are very much random and stochastic in nature. The randomness can only be properly understood through the statistics. Network science uses several powerful statistical tools (Fienberg, 2012). Therefore it is good at the analysis of real world complex problems. Similarly, complex system modeling, analysis and design need the principles of network science for simplification. Network science correlates many diverse fields in a very elegant way. It is now applicable to fields where mathematics had no previous scope. Most of the evolving networks are complex and need advanced tools for their analysis. Complexities

in the networks can be properly modeled through network science. Artificial intelligence can be well complemented through network science in the complex problems. Network science has a great potential to analyze the complex problems of engineering across several disciplines.

Network science is a natural tool for the solutions of the complex engineering problems. For instance, the Internet was developed in the 1970s, but became popular only after the arrival of World Wide Web (WWW), because the network dynamics of a large globally connected network was not known. That is the reason why many engineers and scientists have a lot of doubts on the stability and sustainability of the Internet. Through the WWW framework, the Internet was expanded very fast. It provided both commercial and non-commercial services around the world. Several new economical and philanthropic activities were started through the Internet. However, its principles and statistical mechanism were understood much later in Albert & Barabási (2002). Similarly, the failures in the power systems are quite complex. Mainly the faults like the cascade failure are very much catastrophic for the power systems. Several such failures have taken place in the past in many countries. However, the solutions to those failures were not available to the engineers. That is mainly to understand the power failures in the transmission networks they had to understand the networks first. In fact, there was no such tool to analyze the huge power networks. In contrast, now network science provides proper framework to analyze the power networks (Buldyrev et al., 2010). Similarly, the inland transportations across the world are carried out through the road networks. These networks in large countries are too big for the engineers to analyze them when something like optimization or renovation is required. Using network science based solutions it can be simplified. The global trade networks are really huge. They cannot be studied accurately without network science.

In the above examples, we observed that the complex problems of the modern world is too complicated and needs smart solutions. In order to understand the problems, the engineers need to understand their underlying complexities. We have seen that these problems can be understood using network science. This is how network science has evolved as a natural solution for complex engineering problems in the recent years. In the long term, network science is going to be an integral part of modern engineering to deal with the complex problems.

Network Science in Communication Engineering

Communication engineering is associated with large communication networks. Therefore, it is associated with network science by default. Communication networks are made up of nodes and links. Nodes are the information hubs of networks. They provide basic signal processing and their conditioning for further utilization elsewhere in the network. The links basically carry the signal from one node to the other. The links can be wireless or wired. Almost all the core networks are wired. However, a large part of the access network is wireless. That is due to the popular service provisioning through the mobile networks, WiFi, WiMAX and several other wireless technologies. Both the node and link related parameters of communication networks can be analyzed using network science.

Statistical analysis and modeling of networks is now an integral part of network science and engineering (Fienberg, 2012). In case of communication networks, it can be used for the planning and dimensioning. In communication engineering, engineers focus on the analysis and modeling of both node and link parameters for network planning, dimensioning, cost estimations, design, overhauling, renovation and expansion. Node parameter analysis normally provides the network dynamics and the nature of the networks. Link length parameters of the networks provide the cost associated with the links. In the planning and design stage, the estimations are to be done with partial information because the complete

information is not available (Routray et. al., 2013). These are challenging problems and the estimations have to be accurate and trustworthy. In these early stages of planning, network science based statistical models can be used to estimate the parameters of network economics (Routray et. al., 2017). Estimation of link related parameters of communication networks using the statistical models obtained from the network science is popular and trustworthy (Routray et. al., 2014).

Main motivation for network science in communication engineering comes from its abilities to analyse the networks and its characteristics. Statistical analysis and modeling of the network parameters are essential components of network science and engineering. They provide several critical behavioural aspects of the networks (Korotky, 2004). Network dynamics cannot be understood without the statistical analysis. In communication engineering they are helpful in the estimation of network parameters and characteristics (Korotky, 2004). Network classification and characterization are very much simpler through network science. Even, the network stability and traffic engineering aspects can be analyzed using the statistical analysis of communication networks (Korotky, 2004). Quick estimation of network requirements which is required during the faults and network overhaul can be performed with good accuracy using the statistical models (Korotky, 2004; Labourdette et. al., 2005).

Node Modeling

Nodes are the meeting points of the links. The central nodes have a lot of connections or high nodal degrees (nodal degree of a node is the number of links connected to that node). Nodal degrees of networks can be used for their characterization. Based on the distributions of the nodal degrees networks can be random or scale-free. Similarly, when a network provides easily accessible paths between any node pairs, it is called a small-world network. In the following paragraphs these networks are described briefly.

Random Networks

Random networks are those whose nodal degree distribution follows the Poisson's distribution. It is well known that Poisson's distribution has only one parameter for its presentation. This parameter is also the average value of the distribution. When the average value becomes large, the probability distribution function (PDF) of this distribution takes a bell shape (it is also known as the bell curve) which is similar to the PDF of the normal distribution. Random networks are not very well connected (Routray et. al., 2015a). Initially the Internet was thought to be a random network due to its sparse and irregular connections. However, later measurements confirmed that the Internet is not a random network. In the early days such as 1970s and 1980s, the core optical networks used to be random to a large extent (Waxman, 1988; Pavan et. al., 2010). However, from the recent analysis of the core networks, it is found that now they are shifting towards scale-free nature (Routray et. al., 2015a).

Scale-Free Networks

Scale-free networks are very much different from the random networks. Their nodal degrees follow hyperbolic shapes or the power-law distributions (Adamic & Huberman, 2000; Routray et. al., 2015a). Typically, the distributions of the nodal degrees of these networks have shape similar to that of rectangular hyperbolas (Routray et. al., 2015a). There are several scale free networks found in the real world. The Internet is found to be a scale-free network. These scale-free networks exhibit some specific kind of dynamics and very much stable. The Internet too exhibits these characteristics.

Small World Networks

In these networks, the number of hops or links between any two nodes does not exceed a specific threshold. Typically the number of hops (or links) required between any two randomly chosen nodes varies proportionally with the logarithm of the number of nodes N in the network (Routray et. al., 2015a). Therefore, a distant node in the network can be reached by traveling through a very few number of hops or links. Social networks and the neural networks of the brain are typical small world networks. However, these characteristics are also exhibited in some communication networks.

Link Modeling

Links of the networks are the carriers from one node to the other. Sparse networks have very few links than the maximum possible links. However, the core networks are not sparse. Rather we find more links in core networks. Highly demanded core networks may be well connected as a full mesh network. More than the connectivity, the links play bigger roles in determining the survivability of the networks. If a node is connected with the network through just a single link then the failure of the link will stop the services to that node. Lengths of the links are directly associated with their cost parameters. Using the link lengths we can quantify the capital expenditure (CAPEX) of the links. Link parameters or elements such as the repeaters, amplifiers, performance monitors and several other sensors are directly related with the link lengths. These parameters and their numbers can be estimated from the link lengths. If a network is too large or the link lengths are not known then the statistical models of the link lengths can be used (Routray et al., 2015). One such model has been developed for the core networks which can be found in Routray et al. (2013). The link statistical model developed in (Routray et al., 2015) provides the statistical model for link lengths which can be estimated from the very basic information of the network. It needs just the coverage area and the number of nodes of the network. Knowledge of these two parameters will not only provide the statistical model for link lengths but also provides the details of the components required along the links. These estimations are very much realistic and can be used for network dimensioning. Link related CAPEX and OPEX too can be estimated using this model.

In (Routray et. al., 2013), it is shown that the statistical model for link lengths of the core networks follow general extreme value (GEV) distribution. The parameters of the GEV distribution are estimated from the coverage area and the number of nodes of the network. The associated input parameters of the statistical model such as the coverage area can be found from the network node locations (Routray et. al., 2014a). The developed GEV model is used for the estimation of the link length dependent parameters in core networks (Routray et. al., 2014).

Other Models in Communication Networks

Besides the nodes and links, there are several other parameters in communication networks which play critical roles in its operations. Therefore, the statistical models for the other parameters of communication networks are required and it is possible to develop them. In addition to the link lengths, there are several other length parameters which are used for practical communication engineering. Traditionally, the shortest path lengths are used for traffic routing and minimization of energy consumption in the transmission. However, in the modern networks we also need the statistical models for the minimum hop path lengths. These minimum hop paths may not be the shortest paths, but they encounter the least number of nodes along their journey. Nodes are the main processing points along the communication paths. So, more

number of nodes along a path means the time spent in the intermediate processing is longer. Therefore, the delay in the transmission is proportionately longer. Minimum hop paths provide the guarantee that the nodal processing delays are at their minimum level. Similarly, backup path lengths are required in case of a node or link failure. Restoration of the services along the alternative routes needs the backup paths. In the modern world where 99.999% reliability in service provisioning is the lower threshold these statistical models provide a lot of advantages in traffic planning and resource optimization.

The shortest path lengths for the core communication networks are found follow Johnson S_B distribution (Routray et. al., 2015). The four parameters of this distribution can be estimated from the coverage area and the number of nodes of the network. This model can be applied for the practical estimations of shortest path dependent parameters. The lower bound of energy consumption in transmission can also be estimated using this model.

FUTURE RESEARCH DIRECTIONS

Network science has a very bright future as it is the main tool to analyze the complex systems. There are many complex systems found in the real world. Network science is used in their studies and provides proper analysis. In medicine, it has been found that several diseases are genetic and thus are transferred from one generation to the other genetically. This is a complex field to analyze the genetic links from one generation to the other. However, network science is able to analyze them quite smoothly. This branch of medicine is known as network medicine as the genetic links are established using a network based model (Barabási et al., 2011). Based on the studies of network medicine, a human disease has been developed which can show the potential transferable diseases from one generation to the next. It also shows the links between the group of similar diseases. In the future, this study of network medicine will be further improved. In fact, all the diseases can be explained using some kind of network based models. Thus, network medicine is regarded as one of the main tools for precision health studies. It looks very clear that it will be very popular in the future.

Similarly, the power grids are being interconnected for better management of power generation, and distribution. Even the smart grid initiatives are being considered for better energy efficiency. The interconnections of grids make them very complex and difficult to manage. Now, several network science based methods are being used to study these huge grids. In fact, the nature and dynamics of these grids are not formulated according to the principles of network science. Now, network science is being applied to analyze the power grids. Hopefully, the future failures of the power grids can be prevented using the power system dynamics of the grids.

With the arrival of the new mobile generations and the Internet of things (IoT), the human living environment will be very much networked. IoT and its pervasive applications would make the whole world a very complex network. In order to understand the IoT properly, network science is required. Its effective design, energy efficiency, monitoring and all operational aspects can be analyzed using network science. IoT and the upcoming new mobile generations can have several good and bad effects on the human lives. Therefore, the mobile networks, IoTs and their other supporting networks must be studied carefully. This is clearly a future application network science.

Chaos is considered as a natural phenomenon of the universe. It is found in every corner of the universe at different scales such as in the femto, nano, and micro scales in the materials, and also at a macro scale covering a large portion of the universe. For instance, the movement of the galactic clusters and their constituents show several chaotic behaviors. Similarly, the motion of the micro, and nano-particles too

show a lot of random and chaotic characteristics. These chaotic behaviors can be analyzed accurately through network science. This is also a future application network science. Like network science, artificial intelligence is also a powerful tool to analyze the system behavior. Even both the fields use the behavioral statistics to a large extent. Thus the conjugation of network science and artificial intelligence looks like a natural choice. In fact, these two tools can analyze the complexities better than any other. So, in the future, network science can be a natural companion of artificial intelligence and machine learning in the complex system analysis.

CONCLUSION

Network science is an elegant field of study and analysis of complex systems. Complexities cannot be analyzed properly without network science. It is an interdisciplinary field of science which interconnects many different fields. It provides several creative insights in the field where it is applied and their outcomes help us to simplify the problems in other fields. It is in fact true that understanding one kind of complex network might shed light on others. In communication engineering it can be used as tool for practical estimation and dimensioning. A realistic model of a communication network can provide several insights for practical problems. Therefore, network science will always be treated as a practical tool for addressing the problems of communication engineering and its networks.

REFERENCES

- Adamic, L. A., Huberman, B. A., Barabási, A.-L., Albert, R., Jeong, H., & Bianconi, G. (2000). Power-law distribution of the world wide web. *Science*, 287(5461), 2115–2115. doi:10.1126/science.287.5461.2115a
- Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47–97. doi:10.1103/RevModPhys.74.47
- Alexanderson, G. (2006). About the cover: Euler and Königsberg's Bridges: A historical view. *Bulletin of the American Mathematical Society*, 43(4), 567–573. doi:10.1090/S0273-0979-06-01130-X
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews. Genetics*, 12(1), 56–68. doi:10.1038/nrg2918 PMID:21164525
- Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916), 892–895. doi:10.1126/science.1165821 PMID:19213908
- Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., & Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291), 1025–1028. doi:10.1038/nature08932 PMID:20393559
- Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1), 17–60.
- Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4), 825–839. doi:10.1080/10618600.2012.738106

Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2), 129–233.

Granovetter, M. S. (1977). The strength of weak ties. In *Social networks* (pp. 347–367). Academic Press. doi:10.1016/B978-0-12-442450-0.50025-0

Korotky, S. K. (2004). Network global expectation model: A statistical formalism for quickly quantifying network needs and costs. *Journal of Lightwave Technology*, 22(3), 703–722. doi:10.1109/JLT.2004.825756

Labourdette, J. F., Bouillet, E., Ramamurthy, R., & Akyamaç, A. A. (2005). Fast approximate dimensioning and performance analysis of mesh optical networks. *IEEE/ACM Transactions on Networking*, 13(4), 906–917. doi:10.1109/TNET.2005.852880

Moreno, J. L. (1934). *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and Mental Disease Publishing Company. doi:10.1037/10648-000

Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A*, 374(1), 457–470. doi:10.1016/j.physa.2006.07.017

Newman, M. E., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(suppl 1), 2566–2572. doi:10.1073/pnas.012582999 PMID:11875211

Pavan, C., Morais, R. M., da Rocha, J. R. F., & Pinto, A. N. (2010). Generating realistic optical transport network topologies. *Journal of Optical Communications and Networking*, 2(1), 80–90. doi:10.1364/JOCN.2.000080

Routray, S. K., da Rocha, J. R. F., & Pinto, A. N. (2014a). Estimating the parameters of optical transport networks from their circumferential ellipses. In *Telecommunications (ICT), 2014 21st International Conference on* (pp. 119–123). IEEE. 10.1109/ICT.2014.6845092

Routray, S. K., Javali, A., Jha, M. K., Nyamangoudar, R., & Sarkar, S. (2017). A Network Science Approach to Estimate Link-Related Network Economics. In *Wireless Communications, Signal Processing and Networking (WiSPNET), 2017 4th International Conference on* (pp. 971–975). IEEE. 10.1109/WiSPNET.2017.8299906

Routray, S. K., Morais, R. M., da Rocha, J. R. F., & Pinto, A. N. (2013). Statistical model for link lengths in optical transport networks. *Journal of Optical Communications and Networking*, 5(7), 762–773. doi:10.1364/JOCN.5.000762

Routray, S. K., Pavan, C., Sahin, G., da Rocha, J. R. F., & Pinto, A. N. (2015a). Statistical analysis and modeling in optical transport networks: characterization and application. *Proceedings of conference on telecommunications (ConfTele)*.

Routray, S. K., Sahin, G., da Rocha, J. R., & Pinto, A. N. (2014). Estimation of link-dependent parameters in optical transport networks from statistical models. *Journal of Optical Communications and Networking*, 6(7), 601–609. doi:10.1364/JOCN.6.000601

Routray, S. K., Sahin, G., da Rocha, J. R. F., & Pinto, A. N. (2015). Statistical analysis and modeling of shortest path lengths in optical transport networks. *Journal of Lightwave Technology*, 33(13), 2791–2801. doi:10.1109/JLT.2015.2413674

- Travers, J., & Milgram, S. (1967). The small world problem. *Psychology Today*, 1(1), 61–67.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684), 440.
- Waxman, B. M. (1988). Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9), 1617–1622. doi:10.1109/49.12889

ADDITIONAL READING

- Malladi, M., Mithun, V., Naveen, K. B. N., Punith, N., Roogi, J. M., Sarma, M. J., & Routray, S. K. (2019). Statistical Analysis of Path Length in Optical Networks. In *Innovations in Electronics and Communication Engineering* (pp. 459–466). Springer. doi:10.1007/978-981-10-8204-7_46
- Priyanka, N., Janhavi, S., Subramanya, P., Naveen, L., Nyamangoudar, R., Routray, S. K., & Suganya, J. (2017). Emergence of small world behaviors in optical transport networks. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 783–786). IEEE. 10.1109/ICICI.2017.8365243
- Routray, S. K., & Javali, A. (2016). Optical performance monitor placement in OTNs using link statistical model. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 1016–1019). IEEE. 10.1109/RTEICT.2016.7807983
- Routray, S. K., Mahto, A., Kumar, D. R., Gunesegar, K., & Choudhury, R. (2018). Link Statistical Pattern of Undersea Communication Networks. In *Innovations in Electronics and Communication Engineering* (pp. 407–412). Springer. doi:10.1007/978-981-10-3812-9_43
- Routray, S. K., Thakeran, R. M., Srikanth, D. N., Preeti, P., & Keerthi, K. (2017). Statistical analysis and modeling of honeybee foraging networks for apiculture. In *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)* (pp. 164–167). IEEE. doi:10.1109/TIAR.2017.8273708
- Sahu, A., Kapur, H., Anand, P., & Routray, S. K. (2018). Analysis of Statistical Pattern in Multi-hop Cellular Networks. In *Innovations in Electronics and Communication Engineering* (pp. 437–443). Springer. doi:10.1007/978-981-10-3812-9_46
- Shreesha, S., & Routray, S. K. (2016). Statistical modeling for communication networks. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)* (pp. 831–834). IEEE. 10.1109/ICATCCCT.2016.7912115

KEY TERMS AND DEFINITIONS

Communication Engineering: This is the branch of electrical engineering that deals with the principles and applications of communication processes.

Communication Networks: These are the networks which facilitate the communication between two distant communicators.

Complexity: Complexity is the lack of simplicity in a system. Complex systems need sophisticated tools to analyze them.

Network: A network is a combination of nodes and links. The nodes function as the source or processor of information and the links work as the communicator between the nodes.

Network Cost Estimation: This is a detailed method of estimation of the costs of different types of networks using logical methods.

Network Dimensioning: In network dimensioning, all the requirements of a network such as number of network components, bandwidth, and cost parameters are determined using the basic principles of communication engineering.

Network Science: This is the emerging branch of science which deals with the study and analysis of networks to explain the complexities.

Statistical Modeling: Statistical modeling is the process or combination of processes in which the statistical information play the key roles.

Application of Biogeography-Based Optimization to Antennas and Wireless Communications

Sotirios K. Goudos

 <https://orcid.org/0000-0001-5981-5683>

Aristotle University of Thessaloniki, Greece

INTRODUCTION

The Evolutionary Algorithms (EAs) mimic behaviour of biological entities and they are inspired from Darwinian evolution in nature. The EAs have been extensively studied and applied to several real-world engineering problems. Biogeography-based optimization (BBO) (Simon, 2008) is a recently introduced evolutionary algorithm. BBO is based on mathematical models that describe how species migrate from one island to another, how new species arise, and how species become extinct. The way the problem solution is found is analogous to nature's way of distributing species. In the BBO approach there is a way of sharing information between solutions (Simon, 2008), similar to the other evolutionary algorithms such as GAs, DE, and PSO. This feature makes BBO suitable for the same types of problems that the other algorithms are used for, namely high-dimensional data. Additionally, BBO has some unique features, which are different from those found in the other evolutionary algorithms. For example, quite different from GAs, DE and PSO, the set of the BBO's solutions is maintained from one generation to the next and is improved using the migration model, where the emigration and immigration rates are determined by the fitness of each solution. These differences can make BBO outperform other algorithms (Simon, 2008). BBO has been applied successfully to several real world engineering problems (Ashrafinia, Pareek, Naeem, & Lee, 2011; Bhattacharya & Chattopadhyay, 2010; Boussaïd, Chatterjee, Siarry, & Ahmed-Nacer, 2011; S. K. Goudos et al., 2012; Jamuna & Swarup, 2011; Kankanala, Srivastava, Srivastava, & Schulz; Mandal, Bhattacharya, Tudu, & Chakraborty, 2011; Rathi, Agarwal, Sharma, & Jain, 2011; Silva, dos S Coelho, & Freire, 2010).

The purpose of this chapter is to briefly describe the BBO algorithm and present its application to antenna and wireless communications design problems. This chapter presents results from design cases that include patch antenna, linear antenna array, and a Partial Transmit Sequence (PTS) scheme for OFDM signals based on BBO. The chapter is supported with an adequate number of references. This chapter is subdivided into five sections. The "Background" Section presents the issues, problems, and trends with BBO. Then we briefly present the main BBO algorithm. In the next Section, we describe the design cases and present the numerical results. An outline of future research directions is provided in the following Section while in the "Conclusion" Section we conclude the chapter and discuss the advantages of using a BBO-based approach in the design and optimization of wireless systems and antennas. Finally, an "Additional Reading Section" gives a list of readings to provide the interested reader with useful sources in the field.

DOI: 10.4018/978-1-7998-3479-3.ch066

BACKGROUND

The mathematical models of Biogeography are based on the work of Robert MacArthur and Edward Wilson in the early 1960s. Using this model, they have been able to predict the number of species in a habitat. The habitat is an area that is geographically isolated from other habitats. The geographical areas that are well suited as residences for biological species are said to have a high habitat suitability index (HSI). Therefore, every habitat is characterized by the HSI which depends on factors like rainfall, diversity of vegetation, diversity of topographic features, land area, and temperature. Each of the features that characterize habitability is known as suitability index variables (SIV). The SIVs are the independent variables while HSI is the dependent variable.

Similar to the other evolutionary algorithms (EAs) such as Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) optimization, in the BBO approach there is a way of sharing information between solutions (Simon, 2008). This feature makes BBO suitable for the same types of problems that the other algorithms are used for, namely high-dimensional data. Additionally, BBO has some unique features that are different from those found in the other evolutionary algorithms. For example, quite different from GAs, Ant Colony Optimization (ACO) and PSO, from one generation to the next the set of the BBO's solutions is maintained and improved using the migration model, where the emigration and immigration rates are determined by the fitness of each solution. BBO differs from PSO in the fact that PSO solutions do not change directly; the velocities change. The BBO solutions share directly their attributes using the migration models. The migration operator provides BBO with a good exploitation ability. These differences can make BBO outperform other algorithms (Ma, 2010; Ma, Simon, Fei, & Chen, 2013; Simon, 2008). It must be pointed out if PSO or ABC are constrained to discrete space then the next generation will not necessarily be discrete (Ma et al., 2013). However, this is not true for BBO; if BBO is constrained to a discrete space then the next generation will also be discrete to the same space. As the authors in (Ma et al., 2013) suggest this indicates that BBO could perform better than other EAs on combinatorial optimization problems, which makes BBO suitable for application to the PTS problem. The main computational cost of EAs is in the evaluation of the objective function. The BBO mechanism is simple, like that of PSO and ABC. Therefore, for most problems, the computational cost of BBO and other EAs will be the same since it will be dominated by objective function evaluation (Ma, 2010). More details about the BBO algorithm can be found in (Ma et al., 2013; Simon, 2008).

BIOGEOGRAPHY-BASED OPTIMIZATION ALGORITHM

Therefore, a solution to a D-dimensional problem can be represented as a vector of SIV variables $[SIV_1, SIV_2, \dots, SIV_D]$, which is a habitat or island. The value of HSI of a habitat is the value of the objective function that corresponds to that solution and it is found by

$$HSI = F(\text{habitat}) = F(SIV_1, SIV_2, \dots, SIV_D)$$

Habitats with a high HSI are good solutions of the objective function, while poor solutions are those habitats with a low HSI. The habitats with high HSI are those that have large population and high emigration rate μ . For these habitats, the immigration rate λ is low. The immigration and emigration rates

are functions of the rank of the given candidate solution. The rank of the given candidate solution represents the number of species in a habitat. These are given by

$$\mu_k = E \left(\frac{k}{S_{\max}} \right), \quad \lambda_k = I \left(1 - \frac{k}{S_{\max}} \right)$$

here I is the maximum possible immigration rate, E is the maximum possible emigration rate, k is the rank of the given candidate solution, and $S_{\max}$ is the maximum number of species (e.g. population size). The rank of the given candidate solution or the number of species is obtained by sorting the solutions from most fit to least fit, according to the HSI value (e.g. fitness). That is the best candidate solution has a rank of $S_{\max}$ and the worst candidate solution has a rank of one. BBO uses both mutation and migration operators. The application of these operators to each SIV in each solution is decided probabilistically. In each generation, there is a probability $P_{\text{mod}} \hat{I} [0, 1]$ that each candidate solution will be modified by migration. P_{mod} is a user-defined parameter that is typically set to a value close to one, and is analogous to crossover probability in GAs. That is the migration operator in each generation is applied only if P_{mod} is greater than a random number generated from a uniform distribution from the interval $[0,1)$.

The migration for NP habitats can be described as follows:

Algorithm 1. Habitat migration

- 1: for $i=1$ to NP
- 2: Select with probability based on
- 3: if then
- 4: for $j=1$ to NP
- 5: Select with probability based on
- 6: if then
- 7: Randomly select a SIV from
- 8: Replace a random SIV in with
- 9: end if
- 10: end for
- 11: end if
- 12: end for

The X_i in the above algorithm is habitat i . The information sharing between habitats is accomplished using the immigration and emigration rate. The λ_i is proportional to the probability that SIVs from neighboring habitats will migrate into habitat X_i . The μ_i is proportional to the probability that SIVs from habitat X_i will migrate into neighboring habitats.

The mutation rate m of a solution S is defined to be inversely proportional to the solution probability and it is given by

$$m_S = m_{\max} \left(\frac{1 - P_s}{P_{\max}} \right)$$

where P_s is the probability that a habitat contains S species (e.g. the rank of the solution is S), $P_{\max}$ is the maximum P_s value over all $s \in [1, S_{\max}]$, and $m_{\max}$ is a user-defined parameter.

Mutation can be described as follows:

Algorithm 2. Habitat mutation

- 1: for $i=1$ to NP
- 2: Compute the probability
- 3: Select SIV with probability based on
- 4: if then
- 5: Replace with a randomly generated SIV
- 6: end if
- 7: end for

The m_i in the above algorithm is the mutation rate of solution i . As with other evolutionary algorithms, BBO also incorporates elitism. This is implemented with a user-selected elitism parameter p . This means that the p best solutions remain from one generation to the other.

APPLICATIONS, RESULTS, AND DISCUSSION

In this Section, we provide representative wireless communications design problems that are solved with the BBO algorithm. First, we present an example case of linear array synthesis by reducing the number elements required. The second example is that of an arbitrary-shaped patch antenna which can be used for satellite communications. The third example uses the BBO algorithm for addressing a common problem in wireless communication system, the reduction of peak to average power ratio (PAPR) of transmitted signals.

Reducing the Number of Elements in Linear Arrays Using BBO

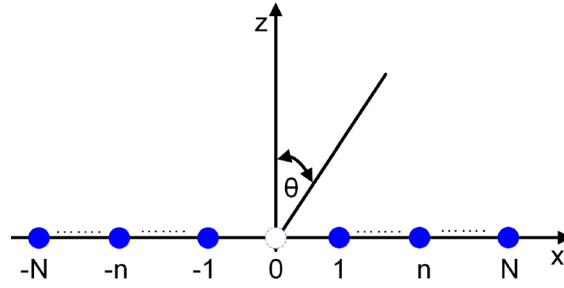
Several applications of modern wireless communications systems require radiation characteristics such as higher directivity that cannot be achieved by using a single antenna. The use of antenna arrays is the obvious solution for such a case. Among others, antenna arrays are employed for several radar and wireless communications applications in space and on earth. Their advantages include the possibility of fast scanning and precise control of the radiation pattern.

Synthesis of linear antenna arrays has been extensively studied in the last decades using several analytical or stochastic methods. The linear array design problem is multimodal and therefore requires the use of an optimization method that does not easily get trapped in local minima. The reduction of array elements plays an important role in cases where the cost, weight and size of antennas are limited. Therefore, a common goal in array synthesis is the generation of a desired pattern using the minimum number of elements. Additional constraints could include the minimum and maximum distance between two adjacent elements.

We assume a $2N+1$ -element linear array, which is symmetrically placed along the x-axis, see Figure 4.

We examine the synthesis of a pattern taken from the literature. We assume that $d_{\min} = 0.5\lambda$, and $d_{\max} = \lambda$, where $d_{\min}, d_{\max}$ is the minimum and maximum distance between two adjacent elements respectively.

Figure 1. Geometry of the antenna array.



The desired pattern in this case is that of a 29-element Taylor-Kaiser array uniformly spaced with $d = 0.5\lambda$, $SLL_d = -25$ dB, and $BW_d = 11.8^\circ$. The authors in (Yanhui, Zaiping, & Qing Huo, 2008) use a non-iterative synthesis method which is based on the matrix pencil method (MPM) (S. Yang, Y. Liu, & Liu, 2009; Yanhui, Qing Huo, & Zaiping, 2010; Yanhui et al., 2008). We set the habitat modification probability, P_{mod} , to 1 and the maximum mutation rate, m_{max} , equal to 0.001. The maximum immigration rate I, and the maximum emigration rate E are both set both to one. The elitism parameter p is set to two. The population size and the number of generations is set to 100 and 2000 respectively.

We run the algorithm for two cases with 17 and 18 elements respectively

The results from the literature are for $N=17$ (Yanhui et al., 2008), and $N=18$ (Wenji, Lian, & Fang, 2011), elements respectively. Therefore, we obtain reconstructed patterns for both cases. The radiation pattern characteristics are reported in Table 1. We notice that BBO found slightly better results than the one in (Yanhui et al., 2008) for the 17-element case. The peak SLL is better while the beamwidth is the same as that of the desired pattern. The value of MSE is also better. For the 18-element case, BBO outperforms (Wenji et al., 2011). The SLL is improved and the beamwidth exactly matches the desired pattern. The MSE is improved by two orders of magnitude compared with (Wenji et al., 2011).

Table 1. Comparative radiation pattern characteristics for the 29-element array design case

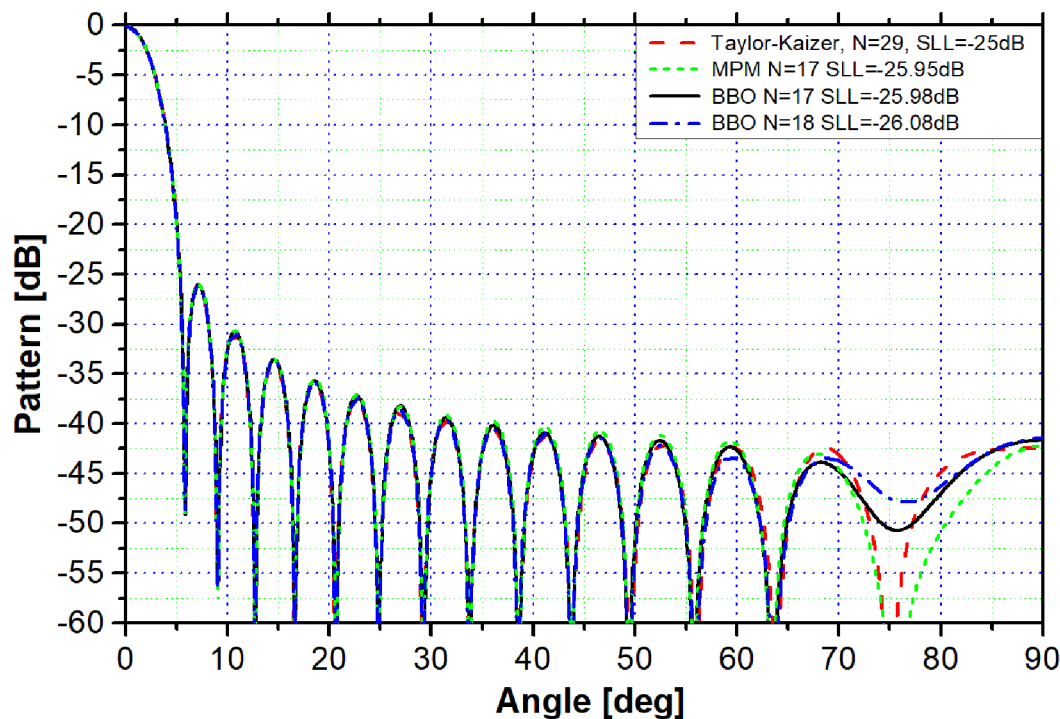
	Desired	MPM (Yanhui et al., 2008) N=17	BBO N=17	MPM (Wenji et al., 2011) N=18	BBO N=18
SLL (dB)	-26.10	-25.96	-25.98	-25.57	-26.04
BW (deg)	11.80	11.80	11.80	12.00	11.80
MSE	n/a	9.586e-7	6.260e-7	2.651e-5	4.192e-7

Arbitrary-shaped Dual Band Patch-Antenna Design Using the BBO Algorithm

The use of microstrip patch antennas in wireless communication systems provides several advantages like low profile, low cost, and ease of fabrication. Patch antennas or arrays of such antennas in the Ku-band have practical applications in radars (Bilgic & Yegin, 2014), and satellite communications (Huayan et al., 2014; Sharma, Yadav, Kumar, Ranga, & Bhatnagar, 2011; Zhang, Mahe, & Razban, 2014).

The next design case is that of arbitrary-shaped dual band patch antenna. Figure 3 shows the patch geometry. It consists of a 5X5 grid where the blue indicates the existence of metal and the white indicates

Figure 2. Synthesis of a 29-element Taylor-Kaiser pattern with $SLL = -25\text{dB}$ with $N= 17$ and $N=18$ elements using BBO



dielectric. The grid is numbered from 1 to 25 as it is shown in the figure. The patch antenna is fabricated on a dielectric substrate with $\epsilon_r = 3.38$ and 30mils thick. Aperture coupled feed is used and an H-shaped slot is created in the ground plane. The design goal is to minimize the S11 magnitude in two frequencies 12GHz and 14GHz. For this case, the population size is set to 10 vectors and the maximum

Figure 3. Arbitrary-shaped patch (a) geometry. The blue boxes indicate presence of metal. (b) best antenna found by BBO

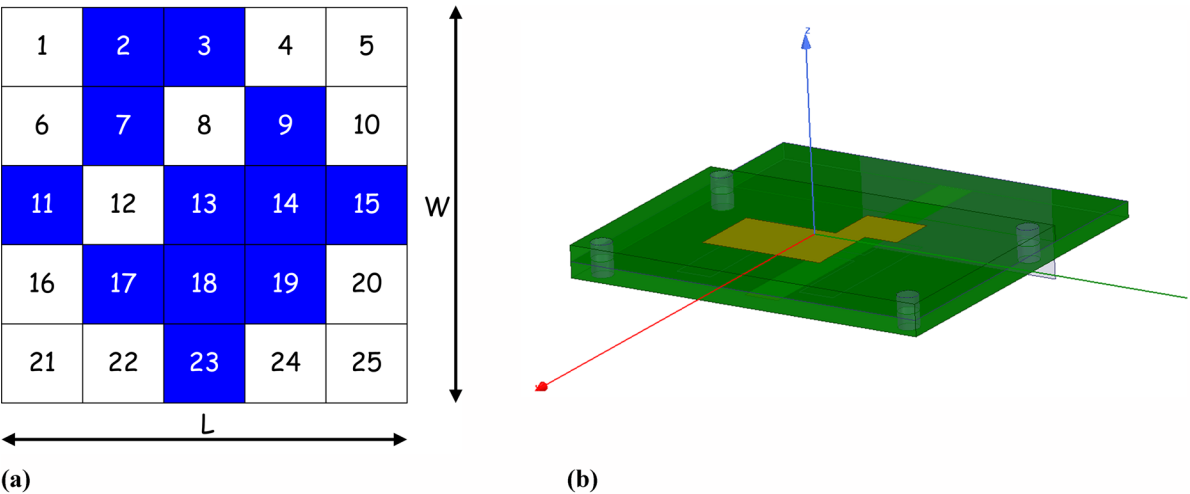


Figure 4. Simulated surface current distribution of the best design found by BBO (a) at 12 GHz, (b) at 14 GHz.

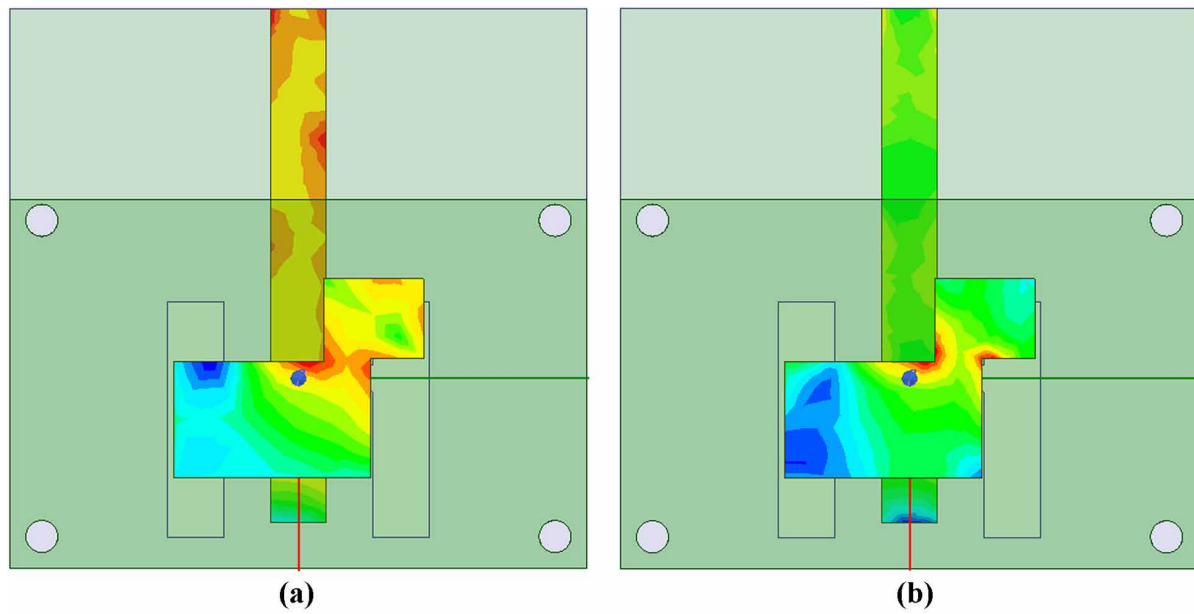
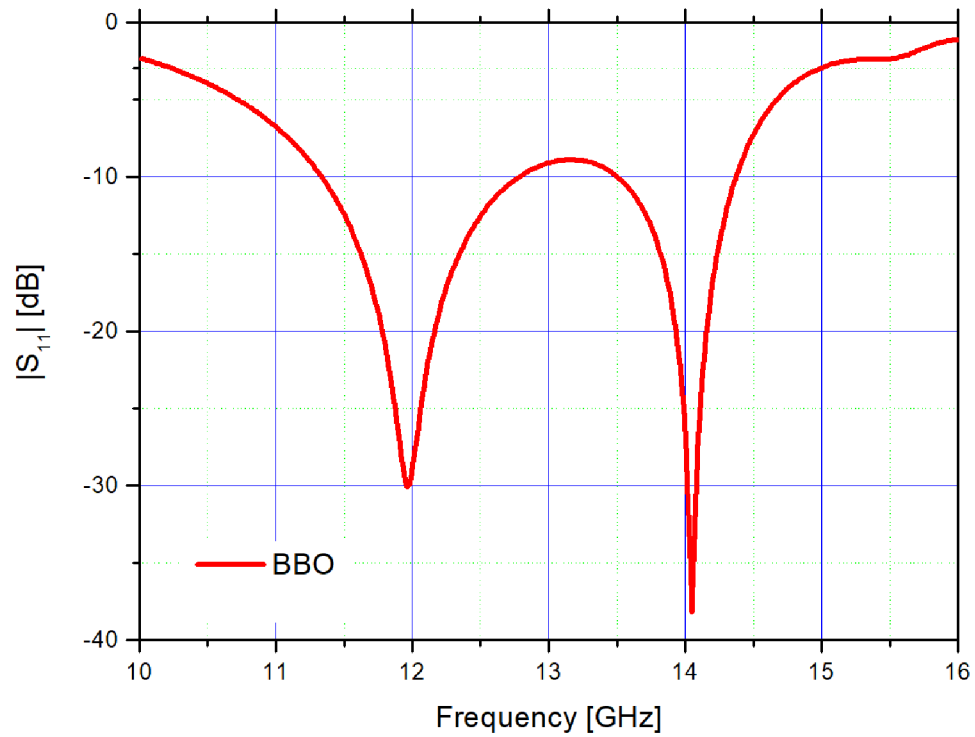


Figure 5. Simulated S_{11} curve for the best design found by BBO.



number of iterations is set to 200. We run the BBO algorithm for 20 independent trials. The surface current distributions for the best-obtained design and the geometry are shown in Figure 4. Figure 5 presents the frequency response of the best design. The BBO obtained antenna has a $|S_{11}| < -10\text{dB}$ bandwidth between 11.35 and 12.78 GHz and from 13.51GHz to 14.37GHz. The S_{11} values at 12 and 14 GHz are found to be -28.45 and -27.59 respectively.

A PTS Technique Based on BBO for PAPR Reduction of OFDM Signals

Orthogonal frequency division multiplexing (OFDM) is widely used in several high-bit-rate digital communication systems such as Digital Audio Broadcasting (DAB), Digital Video broadcasting (DVB) and wireless local area networks (Hara & Prasad, 1997; Van Nee & Prasad, 2000). OFDM systems still have several research challenging issues. A major drawback of OFDM signals is the high value of peak to average power ratio (PAPR). Thus, it is important to reduce the PAPR of OFDM signal in order to fully utilize the OFDM technical features. Partial transmit sequences (PTS) (Cimini Jr & Sollenberger, 2000), is a popular PAPR reduction method with good PAPR reduction performance. However, PTS requires an exhaustive search in order to find the optimal phase factors. Thus, the search complexity is high. Several methods have been published in the literature for PAPR reduction using PTS with low search complexity

In an OFDM system, the high-rate data stream is split into N low-rate data streams that are simultaneously transmitted using N subcarriers. The discrete-time signal of such a system is given by

$$s_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} S_n e^{j \frac{2\pi nk}{LN}} \quad k = 0, 1, \dots, LN - 1$$

where L is the oversampling factor, $\mathbf{S} = [S_0, S_1, \dots, S_{N-1}]^T$ is the input signal block. Each symbol is modulated by either phase-shift keying (PSK) or quadrature amplitude modulation (QAM).

In the PTS approach the $\mathbf{S}$ input data OFDM block is partitioned into M disjointed subblocks represented by the vector $\mathbf{S}_m$ $m = 1, 2, \dots, M - 1$ and oversampled by inserting $(L - 1)N$ zeros. Next, the subblocks are converted to time domain using LN point inverse fast fourier transform (IFFT). The PTS objective is to produce a weighted combination of the M subblocks using $\mathbf{b} = [b_1, b_2, \dots, b_M]^T$ complex phase factors to minimize PAPR. In order to reduce the search complexity the phase factor possible values are limited to a finite set. The set of allowable phase factors is

$$A = \left\{ e^{j \frac{2\pi n}{W}} \mid n = 0, 1, \dots, W - 1 \right\}$$

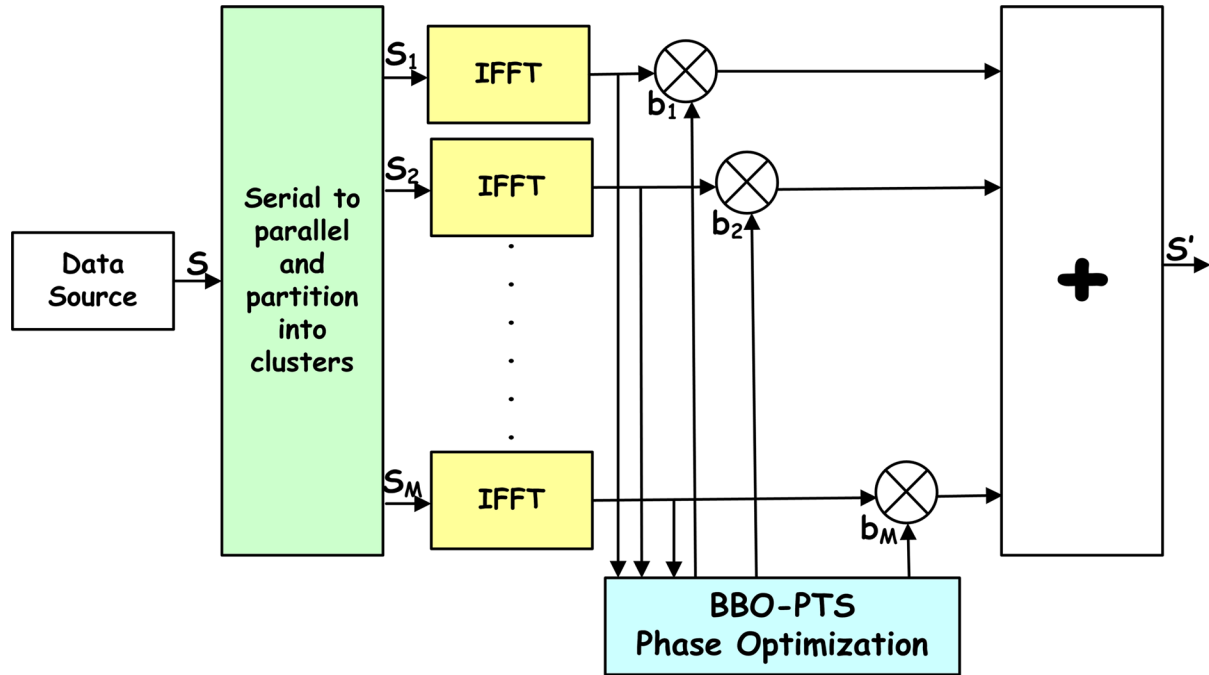
where W is the number of allowed phase factors. Therefore in case of M subblocks and W phase factors the total number of possible combinations is W^M . In order to reduce the search complexity we usually set fixed one phase factor.

The block diagram of the PTS technique is shown in Figure 6.

The optimization goal of the PTS scheme is to find the optimum phase combination for minimum PAPR. Thus, the objective function can be expressed as

Minimize

Figure 6. Block diagram of the PTS technique



$$F(\mathbf{b}) = 10 \log_{10} \frac{\max_{0 \leq k \leq LN-1} |\mathbf{s}'(\mathbf{b})|^2}{E[|\mathbf{s}'(\mathbf{b})|^2]}$$

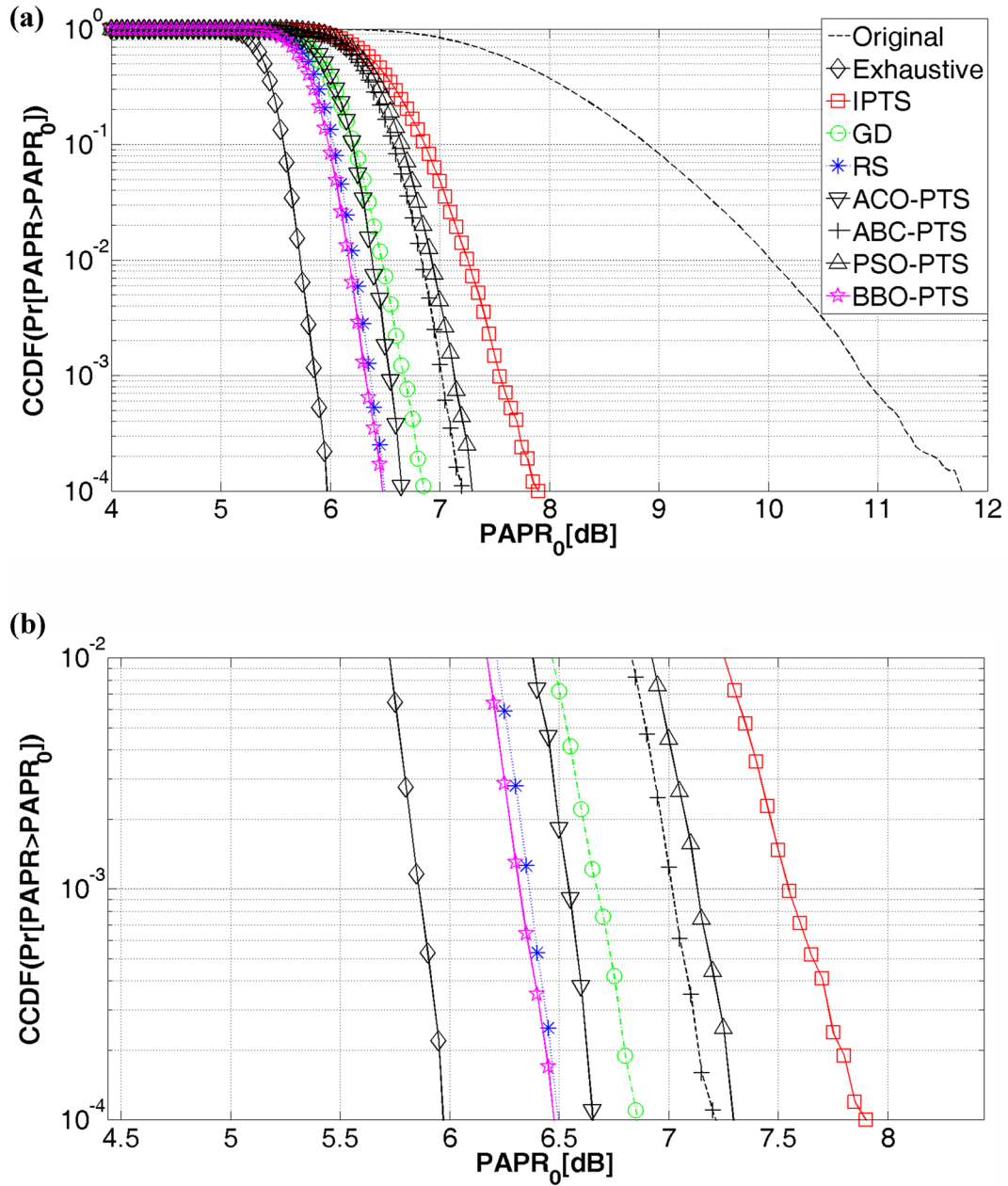
subject to

$$\mathbf{b} \in \left\{ e^{j\phi_m} \right\}^M \text{ where } \phi_m \in \left\{ \frac{2\pi l}{W} \mid l = 0, 1, \dots, W-1 \right\}$$

We have evaluated the objective function above using different evolutionary algorithms and methods found in the literature. We have used two main measurement criteria namely the complementary cumulative distribution function (CCDF) and the computational complexity. In all our simulations, 10E5 random OFDM blocks are generated. The transmitted signal is oversampled by a factor $L=4$. We consider 16-QAM modulation with $N=256$ sub-carriers which are divided into $M=16$ random subblocks. The phase factors for $W=2$ are selected. We consider the first phase factor to be fixed so the total number of unknown phase factors is $M-1$. Figures 7a and 7b show the comparison between the CCDF by BBO-PTS and other PTS reduction techniques. For $\Pr(PAPR > PAPR_0) = 10^{-3}$ the PAPR of the original OFDM transmitted signal is 10.84dB. For all evolutionary algorithms, the population size NP is set to 30 and the maximum number of generations G is set to 30. Thus, the computational complexity of BBO-PTS for this case is $NP \times G = 900$. The PAPR by the BBO-PTS is 6.32dB. The computational complexity of the exhaustive search is $W^{M-1} = 32768$ while the PAPR for this case is 5.86dB. The PAPR by the iterative flipping algorithm for PTS (IPTS) (Cimini Jr & Sollenberger, 2000) is 7.55dB with search complexity $(M-1)W=30$. The PAPR by the gradient descent method (GD) (Han & Lee, 2004) with search

complexity $C_{M-1}^r W^r I = C_{15}^2 2^2 3 = 1260$ is 6.96dB. We also use the Random Search (RS) (Cimini Jr & Solenberger, 2000) method by selecting 900 random phase factors. The PAPR for this case is 6.37dB. The PAPR by Artificial Bee Colony (ABC) PTS (Y. Wang, Chen, & Tellambura, 2010), Particle Swarm Optimization (PSO) PTS (Wen, Lee, Huang, & Hung, 2008), and Ant Colony Optimization (ACO) PTS (Dorigo & Stutzle, 2004), is 7.01dB, 7.13dB, and 6.52dB, respectively.

Figure 7. a) PARP reduction performance comparison of BBO-PTS algorithm with other PTS schemes for $NP=30$, $G=30$. b) Zoom in view of PARP reduction performance comparison of BBO-PTS.



FUTURE RESEARCH DIRECTIONS

The research domain of biogeography-based optimization is growing rapidly. Several extensions of the original algorithm have been emerged. Among others a trend in BBO research is to use different migration models like curve shapes (Ma & Simon, 2011a), and blended migration (Ma & Simon, 2011b). A performance comparison of different migration models is given in (Guo, Wang, & Wu, 2014) and in (Ma, 2010). Other BBO variants proposed include (Ma & Simon, 2011a), blended migration (Ma & Simon, 2011b), oppositional BBO (Roy & Mandal, 2011), real-coded BBO (Gong, Cai, Ling, & Li, 2010), chaotic BBO (Saremi, Mirjalili, & Lewis, 2014), and BBO with a polyphyletic migration operator (Xiong, Shi, & Duan, 2014).

BBO hybrids with other algorithms is also a growing research trend. BBO has been combined with Differential Evolution (DE) (Gong, Cai, & Ling, 2011), with evolutionary strategies (Du, Simon, & Ergezer, 2009), with Harmony Search (HS) algorithm (G. Wang et al., 2013), and local search methods (Feng et al., 2013). The application of multi-objective BBO algorithms (Costa E Silva, Coelho, & Lebensztajn, 2012; Jamuna & Swarup, 2012; Li & Yin, 2013) to antenna and wireless design problems introduces new challenges regarding performance and computational cost. Recently a novel Multi-objective BBO algorithm has been applied to wireless networks design (Sotirios K. Goudos, Plets, Liu, Martens, & Joseph, 2015). Moreover, a chaotic BBO algorithm was applied in (Tegou et al., 2018) for spectrum allocation in cognitive radio networks. The problem of antenna selection in Multiple Input Multiple Output (MIMO) systems is solved in (Fountoukidis et al., 2018) using BBO variants. Additionally, real-coded BBO is applied by the authors in (Paraskevopoulos, Dallas, Siakavara, & Goudos, 2017) for cognitive engine design. Therefore, new research directions have to be explored.

CONCLUSION

In this chapter, we discussed and evaluated the application of BBO optimizers to antenna and wireless communications design problems. BBO is a recently introduced global stochastic optimizer that is based on mathematical models of species migration. Our results show the efficiency of this approach for antenna synthesis. In the design case presented here, the BBO results are better than those found in the literature. The BBO algorithm has been applied to the design of an arbitrary-shaped patch antenna. The obtained best design is quite efficient in both frequencies. Additionally we have used BBO a new low complexity PTS scheme. The comparison with other existing PTS reduction methods shows that BBO-PTS performs better than other methods with the same computational complexity. Simulation results show that BBO-PTS is an efficient and powerful tool for PAPR reduction of OFDM signals.

Recent BBO algorithms and their application to wireless communications problems will be part of our future work. The BBO algorithms can be valuable tools for constrained single or multi-objective optimization problems in antennas and wireless communications.

REFERENCES

Ashrafinia, S., Pareek, U., Naeem, M., & Lee, D. (2011). *Source and Relay Power Selection Using Biogeography-Based Optimization for Cognitive Radio Systems*. Paper presented at the 2011 IEEE Vehicular Technology Conference (VTC Fall).

- Bhattacharya, A., & Chattopadhyay, P. K. (2010). Biogeography-Based Optimization for Different Economic Load Dispatch Problems. *IEEE Transactions on Power Systems*, 25(2), 1064–1077. doi:10.1109/TPWRS.2009.2034525
- Bilgic, M. M., & Yegin, K. (2014). Wideband Offset Slot-Coupled Patch Antenna Array for X/Ku-Band Multimode Radars. *Antennas and Wireless Propagation Letters, IEEE*, 13, 157–160. doi:10.1109/LAWP.2013.2296911
- Boussaïd, I., Chatterjee, A., Siarry, P., & Ahmed-Nacer, M. (2011). Hybridizing biogeography-based optimization with differential evolution for optimal power allocation in wireless sensor networks. *IEEE Transactions on Vehicular Technology*, 60(5), 2347–2353. doi:10.1109/TVT.2011.2151215
- Cimini Jr, L. J., & Sollenberger, N. R. (2000). Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences. *IEEE Communications Letters*, 4(3), 86–88.
- Costa, E., Silva, M. D. A., Coelho, L. D. S., & Lebensztajn, L. (2012). Multiobjective biogeography-based optimization based on predator-prey approach. *IEEE Transactions on Magnetics*, 48(2), 951–954. doi:10.1109/TMAG.2011.2174205
- Dorigo, M., & Stutzle, T. (2004). *Ant Colony Optimization*. The MIT Press. doi:10.7551/mitpress/1290.001.0001
- Du, D., Simon, D., & Ergezer, M. (2009). Biogeography-based optimization combined with evolutionary strategy and immigration refusal. Paper presented at the Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics. 10.1109/ICSMC.2009.5346055
- Feng, Q., Liu, S., Wu, Q., Tang, G., Zhang, H., & Chen, H. (2013). Modified biogeography-based optimization with local search mechanism. *Journal of Applied Mathematics*, 2013, 1–24. Advance online publication. doi:10.1155/2013/960524
- Fountoukidis, K. C., Kalialakis, C., Psannis, K. E., Siakavara, K., Goudos, S. K., Sarigiannidis, P., & Obaidat, M. (2018). MIMO antenna selection using biogeography-based optimization with nonlinear migration models. *International Journal of Communication Systems*, 31(17), e3813. doi:10.1002/dac.3813
- Gong, W., Cai, Z., & Ling, C. X. (2011). DE/BBO: A hybrid differential evolution with biogeography-based optimization for global numerical optimization. *Soft Computing*, 15(4), 645–665. doi:10.1007/00500-010-0591-1
- Gong, W., Cai, Z., Ling, C. X., & Li, H. (2010). A real-coded biogeography-based optimization with mutation. *Applied Mathematics and Computation*, 216(9), 2749–2758. doi:10.1016/j.amc.2010.03.123
- Goudos, S. K., Baltzis, K. B., Siakavara, K., Samaras, T., Vafiadis, E., & Sahalos, J. N. (2012). Reducing the number of elements in linear arrays using biogeography-based optimization. *Proceedings of 6th European Conference on Antennas and Propagation, EuCAP 2012*. 10.1109/EuCAP.2012.6206314
- Goudos, S. K., Plets, D., Liu, N., Martens, L., & Joseph, W. (2015). A multi-objective approach to indoor wireless heterogeneous networks planning based on biogeography-based optimization. *Computer Networks*, 91, 564–576. Advance online publication. doi:10.1016/j.comnet.2015.08.037
- Guo, W., Wang, L., & Wu, Q. (2014). An analysis of the migration rates for biogeography-based optimization. *Information Sciences*, 254, 111–140. doi:10.1016/j.ins.2013.07.018

Han, S. H., & Lee, J. H. (2004). PAPR reduction of OFDM signals using a reduced complexity PTS technique. *IEEE Signal Processing Letters*, 11(11), 887-890.

Hara, S., & Prasad, R. (1997). Overview of multicarrier CDMA. *IEEE Communications Magazine*, 35(12), 126-133. doi:10.1109/35.642841

Huayan, J., Kuo-Sheng, C., Wenquan, C., Chih-Chun, C., Hua-Juan, L., & Quan, X. (2014). Differential-Fed Patch Antenna Arrays With Low Cross Polarization and Wide Bandwidths. *Antennas and Wireless Propagation Letters, IEEE*, 13, 1069-1072. doi:10.1109/LAWP.2014.2328352

Jamuna, K., & Swarup, K. S. (2011). Power system observability Using Biogeography Based Optimization. *International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011)*. 10.1049/cp.2011.0394

Jamuna, K., & Swarup, K. S. (2012). Multi-objective biogeography based optimization for optimal PMU placement. *Applied Soft Computing*, 12(5), 1503-1510. doi:10.1016/j.asoc.2011.12.020

Kankanala, P., Srivastava, S. C., Srivastava, A. K., & Schulz, N. N. Optimal Control of Voltage and Power in a Multi-Zonal MVDC Shipboard Power System. *IEEE Transactions on Power Systems*, 1-9.

Li, X., & Yin, M. (2013). Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Transactions on Nanobioscience*, 12(4), 343-353. doi:10.1109/TNB.2013.2294716 PMID:25003163

Ma, H. (2010). An analysis of the equilibrium of migration models for biogeography-based optimization. *Information Sciences*, 180(18), 3444-3464. doi:10.1016/j.ins.2010.05.035

Ma, H., & Simon, D. (2011a). Analysis of migration models of biogeography-based optimization using Markov theory. *Engineering Applications of Artificial Intelligence*, 24(6), 1052-1060. doi:10.1016/j.engappai.2011.04.012

Ma, H., & Simon, D. (2011b). Blended biogeography-based optimization for constrained optimization. *Engineering Applications of Artificial Intelligence*, 24(3), 517-525. doi:10.1016/j.engappai.2010.08.005

Ma, H., Simon, D., Fei, M., & Chen, Z. (2013). On the equivalences and differences of evolutionary algorithms. *Engineering Applications of Artificial Intelligence*, 26(10), 2397-2407. doi:10.1016/j.engappai.2013.05.002

Mandal, K. K., Bhattacharya, B., Tudu, B., & Chakraborty, N. (2011). *A novel population-based optimization algorithm for optimal distribution capacitor planning*. Paper presented at the 2011 International Conference on Energy, Automation, and Signal (ICEAS). 10.1109/ICEAS.2011.6147075

Paraskevopoulos, A., Dallas, P. I., Siakavara, K., & Goudos, S. K. (2017). Cognitive Radio Engine Design for IoT Using Real-Coded Biogeography-Based Optimization and Fuzzy Decision Making. *Wireless Personal Communications*, 97(2), 1813-1833. doi:10.1007/11277-017-4646-9

Rathi, A., Agarwal, A., Sharma, A., & Jain, P. (2011). *A new hybrid technique for solution of economic load dispatch problems based on Biogeography Based Optimization*. Paper presented at the TENCON 2011 - 2011 IEEE Region 10 Conference.

- Roy, P. K., & Mandal, D. (2011). Quasi-oppositional biogeography-based optimization for multi-objective optimal power flow. *Electric Power Components and Systems*, 40(2), 236–256. doi:10.1080/15325008.2011.629337
- Saremi, S., Mirjalili, S., & Lewis, A. (2014). Biogeography-based optimisation with chaos. *Neural Computing & Applications*, 25(5), 1077–1097. doi:10.100700521-014-1597-x
- Sharma, M. M., Yadav, S., Kumar, A., Ranga, Y., & Bhatnagar, D. (2011). *Compact elliptical microstrip patch antenna with slotted ground for Ku-band applications*. Paper presented at the Applied Electromagnetics Conference (AEMC).
- Silva, M. A. C., dos S Coelho, L., & Freire, R. Z. (2010). *Biogeography-based Optimization approach based on Predator-Prey concepts applied to path planning of 3-DOF robot manipulator*. Paper presented at the 2010 IEEE Conference on Emerging Technologies and Factory Automation (ETFA). 10.1109/ETFA.2010.5641274
- Simon, D. (2008). Biogeography-Based Optimization. *IEEE Transactions on Evolutionary Computation*, 12(6), 702–713. doi:10.1109/TEVC.2008.919004
- Tegou, T., Tsiflikiotis, A., Siakavara, K., Nikolaidis, S., Goudos, S., Dimitrios, V., ... Obaidat, M. (2018). Spectrum Allocation in Cognitive Radio Networks using Chaotic Biogeography-based Optimization. *IET Networks*. <https://digital-library.theiet.org/content/journals/10.1049/iet-net.2017.0264>
- Van Nee, R., & Prasad, R. (2000). *OFDM for wireless multimedia communications*. OFDM for Wireless Multimedia Communications.
- Wang, G., Guo, L., Duan, H., Wang, H., Liu, L., & Shao, M. (2013). Hybridizing harmony search with biogeography based optimization for global numerical optimization. *Journal of Computational and Theoretical Nanoscience*, 10(10), 2312–2322. doi:10.1166/jctn.2013.3207
- Wang, Y., Chen, W., & Tellambura, C. (2010). A PAPR reduction method based on artificial bee colony algorithm for OFDM signals. *IEEE Transactions on Wireless Communications*, 9(10), 2994–2999. doi:10.1109/TWC.2010.081610.100047
- Wen, J. H., Lee, S. H., Huang, Y. F., & Hung, H. L. (2008). A suboptimal PTS algorithm based on particle swarm optimization for PAPR reduction in OFDM systems. *EURASIP J. Wireless Commun. Networking*, 2008(14).
- Wenji, Z., Lian, L., & Fang, L. (2011). Reducing the Number of Elements in Linear and Planar Antenna Arrays With Sparseness Constrained Optimization. *IEEE Transactions on Antennas and Propagation*, 59(8), 3106–3111. doi:10.1109/TAP.2011.2158943
- Xiong, G., Shi, D., & Duan, X. (2014). Enhancing the performance of biogeography-based optimization using polyphyletic migration operator and orthogonal learning. *Computers & Operations Research*, 41(1), 125–139. doi:10.1016/j.cor.2013.07.021
- Yang, S., Liu, Y., & Liu, Q. H. (2009). Combined strategies based on matrix pencil method and tabu search algorithm to minimize elements of non-uniform antenna array. *Progress in Electromagnetics Research B. Pier B*, 18, 259–277. doi:10.2528/PIERB09092408

Yanhui, L., Qing Huo, L., & Zaiping, N. (2010). Reducing the Number of Elements in the Synthesis of Shaped-Beam Patterns by the Forward-Backward Matrix Pencil Method. *IEEE Transactions on Antennas and Propagation*, 58(2), 604–608. doi:10.1109/TAP.2009.2037709

Yanhui, L., Zaiping, N., & Qing Huo, L. (2008). Reducing the Number of Elements in a Linear Antenna Array by the Matrix Pencil Method. *IEEE Transactions on Antennas and Propagation*, 56(9), 2955–2962. doi:10.1109/TAP.2008.928801

Zhang, H., Mahe, Y., & Razban, T. (2014). Low-cost Ku-band dual-polarized and beam switchable cross-type antenna array for satellite communications. *Microwave and Optical Technology Letters*, 56(11), 2656–2659. doi:10.1002/mop.28670

ADDITIONAL READING

Ashrafinia, S., Pareek, U., Naeem, M., & Lee, D. (2011a, 11-15 April 2011). Biogeography-based optimization for joint relay assignment and power allocation in cognitive radio systems. Paper presented at the 2011 IEEE Symposium on Swarm Intelligence (SIS). 10.1109/SIS.2011.5952589

Ashrafinia, S., Pareek, U., Naeem, M., & Lee, D. (2011b, 5-8 Sept. 2011). Source and Relay Power Selection Using Biogeography-Based Optimization for Cognitive Radio Systems. Paper presented at the 2011 IEEE Vehicular Technology Conference (VTC Fall).

Bhattacharya, A., & Chattopadhyay, P. K. (2010a). Biogeography-Based Optimization for Different Economic Load Dispatch Problems. *IEEE Transactions on Power Systems*, 25(2), 1064–1077. doi:10.1109/TPWRS.2009.2034525

Bhattacharya, A., & Chattopadhyay, P. K. (2010b). Solving complex economic load dispatch problems using biogeography-based optimization. *Expert Systems with Applications*, 37(5), 3605–3615. doi:10.1016/j.eswa.2009.10.031

Bhattacharya, A., & Chattopadhyay, P. K. (2011). Hybrid differential evolution with biogeography-based optimization algorithm for solution of economic emission load dispatch problems. *Expert Systems with Applications*, 38(11), 14001–14010. doi:10.1016/j.eswa.2011.04.208

Boussaïd, I., Chatterjee, A., Siarry, P., & Ahmed-Nacer, M. (2011). Hybridizing biogeography-based optimization with differential evolution for optimal power allocation in wireless sensor networks. *IEEE Transactions on Vehicular Technology*, 60(5), 2347–2353. doi:10.1109/TVT.2011.2151215

Garg, V., & Deep, K. (2015). *Advances in Intelligent Systems and Computing* (Vol. 336). A state-of-the-art review of biogeography-based optimization. Springer Verlag.

Jamuna, K., & Swarup, K. S. (2011, 20-22 July 2011). Power system observability Using Biogeography Based Optimization. Paper presented at the International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011). 10.1049/cp.2011.0394

Lohokare, M. R., Pattnaik, S. S., Panigrahi, B. K., & Das, S. (2013). Accelerated biogeography-based optimization with neighborhood search for optimization. *Applied Soft Computing*, 13(5), 2318–2342. doi:10.1016/j.asoc.2013.01.020

- Pattnaik, S. S., Lohokare, M. R., & Devi, S. (2010). Enhanced biogeography-based optimization using modified clear duplicate operator. Paper presented at the Proceedings - 2010 2nd World Congress on Nature and Biologically Inspired Computing, NaBIC 2010. 10.1109/NABIC.2010.5716322
- Rathi, A., Agarwal, A., Sharma, A., & Jain, P. (2011, 21-24 Nov. 2011). A new hybrid technique for solution of economic load dispatch problems based on Biogeography Based Optimization. Paper presented at the TENCON 2011 - 2011 IEEE Region 10 Conference.
- Roy, P. K., Ghoshal, S. P., & Thakur, S. S. (2010). Biogeography based optimization for multi-constraint optimal power flow with emission and non-smooth cost function. *Expert Systems with Applications*, 37(12), 8221–8228. doi:10.1016/j.eswa.2010.05.064
- Roy, P. K., & Mandal, D. (2011). Quasi-oppositional biogeography-based optimization for multi-objective optimal power flow. *Electric Power Components and Systems*, 40(2), 236–256. doi:10.1080/15325008.2011.629337
- Sharaqa, A., & Dib, N. (2011, 6-8 Dec. 2011). Design of linear and circular antenna arrays using biogeography based optimization. Paper presented at the 2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). 10.1109/AEECT.2011.6132503
- Sharaqa, A., & Dib, N. (2014). Design of Linear and Elliptical Antenna Arrays Using Biogeography Based Optimization. *Arabian Journal for Science and Engineering*, 39(4), 2929–2939. doi:10.1007/13369-013-0794-8
- Silva, M. A. C., dos S Coelho, L., & Freire, R. Z. (2010, 13-16 Sept. 2010). Biogeography-based Optimization approach based on Predator-Prey concepts applied to path planning of 3-DOF robot manipulator. Paper presented at the 2010 IEEE Conference on Emerging Technologies and Factory Automation (ETFA). 10.1109/ETFA.2010.5641274
- Simon, D., Omran, M. G. H., & Clerc, M. (2014). Linearized biogeography-based optimization with re-initialization and local search. *Information Sciences*, 267, 140–157. doi:10.1016/j.ins.2013.12.048
- Singh, U., Kumar, H., & Kamal, T. S. (2010). Design of Yagi-Uda Antenna Using Biogeography Based Optimization. *IEEE Transactions on Antennas and Propagation*, 58(10), 3375–3379. doi:10.1109/TAP.2010.2055778
- Singh, U., Kumar, H., & Kamal, T. S. (2010). Linear array synthesis using biogeography based optimization. *Progress in Electromagnetics Research M. Pier M*, 11, 25–36. doi:10.2528/PIERM09120201
- Wang, L., & Xu, Y. (2011). An effective hybrid biogeography-based optimization algorithm for parameter estimation of chaotic systems. *Expert Systems with Applications*, 38(12), 15103–15109. doi:10.1016/j.eswa.2011.05.011
- Xu, Z. D., Mo, H. W., Xu, Z. J., & Zhang, J. (2013). *Multi-objective biogeography-based optimization algorithm for the control of robotic manipulator* (Vol. 659).

KEY TERMS AND DEFINITIONS

Digital Audio Broadcasting (DAB): A digital radio technology for broadcasting radio stations.

Digital Video Broadcasting (DVB): A suite of internationally accepted open standards for digital television.

Genetic Algorithms: A stochastic population-based global optimization technique that mimics the process of natural evolution.

Mainlobe Beamwidth: In a radio antenna's radiation pattern, the main lobe, or main beam is the lobe containing the maximum power. The beamwidth of the main lobe is the width of the main lobe specified by the angles between the points on the side of the lobe where the power has fallen to zero.

Orthogonal Frequency-Division Multiplexing (OFDM): A method of encoding digital data on multiple carrier frequencies.

Sidelobe Level (SLL): The ratio, usually expressed in decibels (dB), of the amplitude at the peak of the main lobe to the amplitude at the peak of a side lobe.

Wireless Local Area Network (WLAN): A network in which a mobile user can connect to a local area network (LAN) through a wireless (radio) connection. Most modern WLANs are based on IEEE 802.11 standards, marketed under the Wi-Fi brand name.

Bits'-Carrying Capacities of Switched Local Area Networks

7

Monday Ofori Eyinagho <https://orcid.org/0000-0002-5270-3096>*Afe Babalola University, Nigeria***Samuel Oluwole Falaki***Afe Babalola University, Nigeria*

INTRODUCTION

All switched Local Area Networks (switched LANs) have a common parameter, known as 'bits-carrying capacity', which is one of the parameters that network managers and administrators use in managing the performances of the networks. In fact, every network has a limited bits'-carrying capacity; and, an increase in a network's response time, which is a measure of the time between when a user requests a service and the time when the service is granted, is an indication that a network is operating above its bits'-carrying capacity (Forouzan, 2008, p. 876). Knowing the bits'-carrying capacity of any network is therefore, very important with respect to operating/managing the network optimally; that is, the ability to operate the network at all times, at, or near, its bits'-carrying capacity. This is because, if a network is operated above its bits'-carrying capacity, its response time increases, leading to a fall in throughput: on the other hand, it is a waste of resources, to operate a network below its bits'-carrying capacity. However, it appears that there is no available, formal method, or formula in literature for determining this parameter. For example, in Chapter 1 of CISCO CCNA Exploration Companion Guide (2018, p. 15), the following advice was offered to LAN Designers/Installers: When selecting switch hardware for a switched LAN, it is necessary to determine which switches are needed in the core, distribution, and access layers - the three-layer hierarchical network design model enunciated by CISCO CCNA Exploration Companion Guide (2018, p. 2) to accommodate the bandwidth requirement of the LAN. The pertinent question that was not answered in this CISCO documentation, and, has not been answered elsewhere is: How should the bandwidth requirement of a LAN be determined? In this Chapter, the derivation of a formula which can be used to obtain the bits' - carrying capacity (usually referred to as bandwidth by some professionals) of any switched LAN is reported. Also reported, is a formula that can be used to obtain the capacity (that is, bits' capacity, in units, such as, Megabits; Gigabits) to subscribe for, whenever a switched LAN needed to be connected to the network of a service provider, such as connecting the LAN to the Internet via an ISP (Internet Service Provider). Illustrative examples applying both formulas are also given.

BACKGROUND

A common networks' management metric that should always be monitored by the application programs which are used to manage the operations of computer networks is the bits'-carrying capacities. Every

DOI: 10.4018/978-1-7998-3479-3.ch067

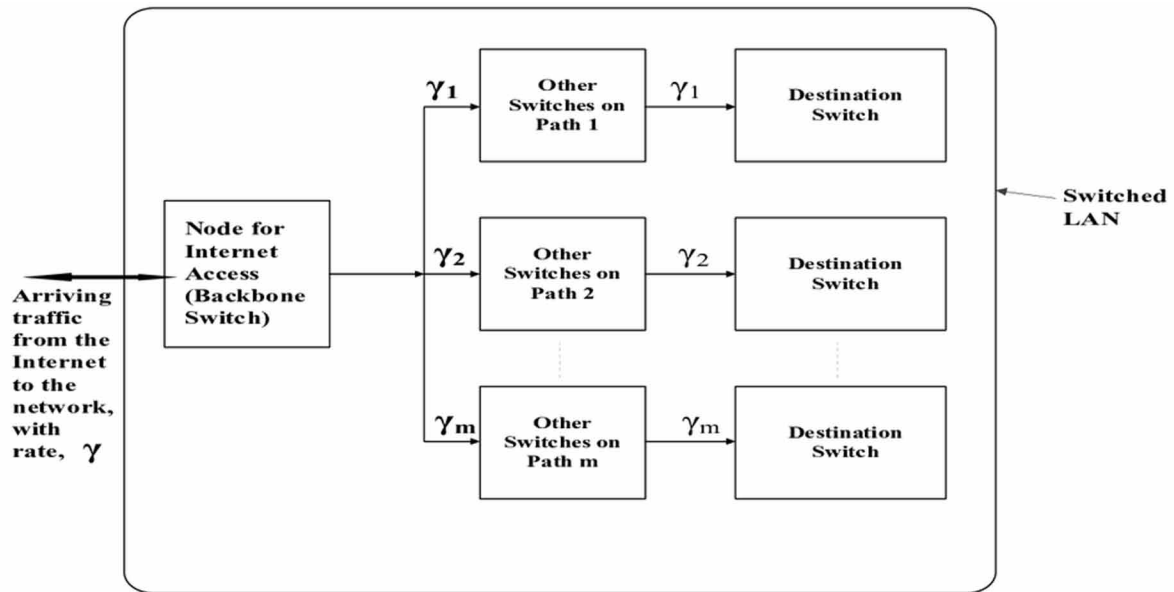
network has a limited bits'-carrying capacity, and one of the functions, according to Forouzan (2008, p. 876), of a network management system (software) is ensuring that a network is not operated above its bits'-carrying capacity (see also, Rathi, N. Malik, Chahal, & S. Malik, 2014; Zubair, Ahmad, Faraz Ahmad, & Ali, 2014).

Bandwidth-Delay Product (BDP) is the term given to the bits'-carrying capacity of a single link (or bits' pipe) in a network; and, is a quantity obtained by multiplying the bandwidth of the link in bits/second, with the round-trip time (RTT) in seconds (that is, the propagation-delay of the link multiplied by 2). The BDP of a link is a very useful quantity to keep in mind when analyzing the performance of computer-communication networks; because, the BDP as defined by Forouzan (2008, p. 93), is the capacity of the bit-pipe from a transmitter to a receiver, and back to the transmitter (see also, Szymanski, 2016; Rangaswamy & Krishnareddy, 2015; Tanenbaum, 2006, p.559). As an example, if the BDP of a bit-pipe is 60 million bits, a transmitter at one end of the pipe would need to transmit a 60 million bits burst to be able to keep transmitting at maximum rate, until it receives the first acknowledgement: in order words, this is the number of bits it takes to fill the pipe. (Tanenbaum, 2006, p.559) In the words of Kai et al. (2004), BDP-UB (Bandwidth Delay Product – Upper Bound) is a path's upper-bound BDP, and this according to the paper, can be considered as a path's maximum packets (bits) carrying capacity. Attempting to operate a network above its BDP-UB does not lead to the achievement of additional throughput; instead, a fall in throughput occurs when more bits than the BDP-UB are injected into the network. (Kai et al. 2004; Gerd, 1989, p.203) Therefore, it is apparent that, knowing the bits'-carrying capacity of any network is very important for the optimal operation of the network. Comer (2004, p.247) in this context, has averred that network administrators and managers are generally agreed on the position that, networks should not be operated above 90% of their bits'-carrying capacities.

FOCUS OF THE ARTICLE

Deriving from the points-of-views that were highlighted in the previous sections, the following main question becomes pertinent: How therefore, should the bits'-carrying capacities of switched LANs be determined? Switched LANs to the best of the knowledge of the research team involved in this work are just arbitrarily installed, without the capacities being known, and, various ad-hoc approaches are usually used for the network management, using network management application programs as the main aids. The research team involved in this work has also not seen it reported in literature, or documented elsewhere of an approach, or, approaches for the determination of the bits'-carrying capacities of switched LANs. Information and Communication Technology (ICT) Managers/Administrators of the Networks of organizations, and, switched LANs Designers/Installers, also appear not to have formal basis for the determination of the bits' capacities (which some erroneously refer to as bandwidths – erroneous because, the unit of interest in this context, is bits, and not, bits/second), which they should advise the management of the organizations they are working for to purchase from Service Providers like ISPs, for Internet access purposes. The capacity (in bits) requirement of any organization for the purpose of Internet access is unarguably, a function of the capacity of the installed base of its LAN – that is, how large the LAN is in terms of the number of its connected end users, which in turn, determines the number of switches that would be installed in the LAN. These ideas are illustrated in Figure 1. From experience subscriptions for bits' capacities has largely been done arbitrarily, using, rule-of-thumb approaches in most cases. The need to provide solutions to these problems was the motivation for the work reported in this article.

Figure 1.



One of the objectives of embarking on researches in this area, which has produced as one of its results, the formulas that are hereunder reported in this article, is the development of methodologies and algorithms, which are of practical utility for solving the performance problems of data communication networks generally, and switched LANs in particular. This stems from the almost non availability of methodologies and algorithms/formulas that can be practically utilized for solving problems associated with this class of networks. In this context, Tanenbaum (2006, p.557) once observed that, understanding network performance is unfortunately more of an art than a science; that, there is little underlying theory which is actually of any use in practice; and that, the best researchers and practitioners can do in this situation is, give rules-of-thumb gained from hard experience, and, present examples taken from the real world. Therefore, in attempting to meet the foregoing objective, this work sought to promote the merger of some level of mathematical elegance, with techniques that are of practical importance.

SOLUTIONS AND RECOMMENDATIONS

This section of the article reports the derivation of formulas that can be used to: 1) calculate the bits'-carrying capacity of any switched LAN, 2) calculate the capacity in bits, for example, Mbits, Gbits, that should be subscribed for (or bought), when it is necessary to connect a switched LAN to the network of an ISP, for the purpose of providing Internet access services to the organization that owns the switched LAN installation. Through an illustrative, real-world example, the need to obtain optimal values of bits'-carrying capacities, and, capacities in bits to be subscribed for, or bought, were recommended.

Derivation of a Formula for the Computation of the Bits'-Carrying Capacities of Switched LANs

Eyinagho et al. (2011), have shown that any switched LAN's end-to-end delays should be enumerated with respect to the switches (nodes) of which the LAN is composed; and, also posited that it is wrong to do the enumeration with respect to the users' hosts (end devices) that are attached to the LAN. How this enumeration should be carried out was illustrated in (Eyinagho, 2012, pp. 24-30): furthermore, the development of a generic maximum delay model of a packet switch, which can be used to compute all the maximum end-to-end delays of any switched LAN, was reported in (Eyinagho, Falaki, & Atayero, 2012, pp. 76-94). According to Metro-Ethernet-Forum (2016), for any bits'-pipe (or link) in a data communication network, bandwidth (BW) is given as:

$$BW \leq \frac{Capacity}{Round-Trip\ Time} \Rightarrow Round-Trip\ Time \times BW \leq Capacity \text{ or } Capacity \geq Round-Trip\ Time (RTT) \times BW \quad (1)$$

Here, BW is the bits'-pipe or communication link's transmission rate (in bits/second); RTT is the time (in seconds) taken for a bit to travel from one end of the pipe or link to the other end, and back; while, Capacity, is the number of bits 'in flight' across the pipe. Stated more succinctly in the context of (1), the Capacity of the pipe refers to the maximum number of bits that can be in transit across it in the time interval, RTT. The implication of (1) is that, lower bound on Capacity = upper-bound on (RTT × BW). However, BW, which is the pipe or link's transmission rate, is a fixed quantity. It therefore, follows that (1) can be restated thus:

$$\text{minimum Capacity} = \text{maximum RTT} \times BW \text{ (a constant)} \quad (2)$$

which can be put simply as:

$$Capacity = BW \times RTT = \text{bandwidth-delay product (BDP)} \quad (3)$$

Capacity in (3) is understood from (2) to be the minimum bits'-carrying capacity that is desired for a link or bits'-pipe; while RTT is the pipe or link's traversing bit, maximum round-trip time. What has generally been reported in literature for the determination of the bits'-carrying capacity of a single, bits'-pipe or link in switched networks is (3); see, for example (Szymanski, 2016; Forouzan, 2008, p. 93; Ryousei, Yuetsu, Tomohiro, Motohico, & Fumihiro, 2006; Tanenbaum, 2006, p. 559; Kai, Yuan, Samarth, & Klara, 2004). This equation is called, 'bandwidth-delay product' (BDP). Switched LANs however, generally consists of several, bits'-pipes or links, interconnected by various switches to form several origin-destination paths. Assume it is desired to have such a LAN, with n switches installed, or desired to be installed, based on the number of users' hosts to be interconnected by the LAN, and the locations of these hosts. Eyinagho, Falaki, and Atayero (2012, p. 134) proposed (4), called the Eyinagho-Falaki maximum end-to-end delay matrix for switched LANs; and, it is a generic maximum end-to-end delay matrix for this class of LANs. With respect to (4), n = the number of switches in a switched LAN; $meted_{ir}$ = maximum end-to-end delay between any host that is connected to switch i and any other host that is connected to switch r; $meted_{ll}$ = maximum end-to-end delay between any two

hosts that are connected to switch 1; $meted_{22}$ = maximum end-to-end delay between any two hosts that are connected to switch 2; ...;

$$\begin{array}{c}
 \begin{array}{cccc}
 & switch1 & switch2 & \dots & switch(n-1) & switch n \\
 switch1 & \left[\begin{array}{ccccc}
 meted_{11} & & & & \\
 meted_{21} & meted_{22} & & & \\
 \vdots & \vdots & \vdots & \vdots & \vdots \\
 meted_{(n-1)1} & meted_{(n-1)2} & \dots & meted_{(n-1)(n-1)} & \\
 meted_{n1} & meted_{n2} & \dots & meted_{n(n-1)} & meted_{nn}
 \end{array} \right]
 \end{array}
 \end{array} \quad (4)$$

$meted_{nn}$ = maximum end-to-end delay between any two hosts that are connected to switch n respectively. If M_{ij} = maximum end-to-end delay of the path connecting origin switch i to destination switch j;

μ_{ij} = total traffic of a subset of the hosts that are connected to switch i to a subset of the hosts that are connected to switch $j \leq$ bandwidth of the path connecting origin switch i to destination switch j (5)

That is, $M_n = (M_{ij})$, and $\mu_n = (\mu_{ij})$, $i=1,2,3,\dots,n$; $j=1,2,3,\dots,n$, is $n \times n$ end-to-end delay matrix, and $n \times n$ bandwidth matrix respectively. It will be shown in this paper that, the capacity of any switched LAN is given by $Tra(M_n \mu_n) = Tra(\mu_n M_n)$; where, $Tra(AB)$ is the trace of the product of matrix **A** and **B**. Therefore,

$$M_n \times \mu_n = \begin{bmatrix} M_{11} & M_{12} & \dots & M_{1n-1} & M_{1n} \\ M_{21} & M_{22} & \dots & M_{2n-1} & M_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ M_{n-11} & M_{n-12} & \dots & M_{n-1n-1} & M_{n-1n} \\ M_{n1} & M_{n2} & \dots & M_{nn-1} & M_{nn} \end{bmatrix} \times \begin{bmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n-1} & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n-1} & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n-11} & \mu_{n-12} & \dots & \mu_{n-1n-1} & \mu_{n-1n} \\ \mu_{n1} & \mu_{n2} & \dots & \mu_{nn-1} & \mu_{nn} \end{bmatrix} \quad (6)$$

The matrix M_n is (4) completed with elements above the leading diagonal; because, elements above and below the leading diagonal are mirror images of each other. Meanwhile, let it be assumed that a hypothetical switched LAN has 2 switches; (6) will become,

$$M_2 \times \mu_2 = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \times \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} = \begin{bmatrix} M_{11}\mu_{11} + M_{12}\mu_{21} & M_{11}\mu_{12} + M_{12}\mu_{22} \\ M_{21}\mu_{11} + M_{22}\mu_{21} & M_{21}\mu_{12} + M_{22}\mu_{22} \end{bmatrix}$$

$$Tra(M_2 \mu_2) = M_{11}\mu_{11} + M_{12}\mu_{21} + M_{21}\mu_{12} + M_{22}\mu_{22} \quad (7)$$

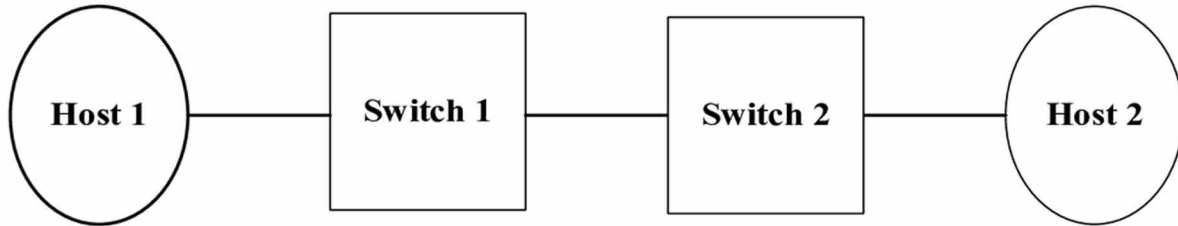
The maximum end-to-end delay of a path in the direction $i \rightarrow j$ is however, the same as the maximum end-to-end delay of the path in the direction $j \rightarrow i$, therefore, $M_{ij} = M_{ji}$; for example, $M_{12} = M_{21}$. It is also, easy to see that the upper bound in (5) is the bandwidth of the path connecting origin switch i to destination switch j, therefore, $\mu_{ij} = \mu_{ji}$. Consequently, replacing M_{ij} with M_{ji} and μ_{ij} with μ_{ji} in (7), gives,

$$\text{Tra}(\mathbf{M}_2\boldsymbol{\mu}_2) = M_{11}\mu_{11} + M_{12}\mu_{12} + M_{12}\mu_{12} + M_{22}\mu_{22}$$

$$\text{Tra}(\mathbf{M}_2\boldsymbol{\mu}_2) = M_{11}\mu_{11} + 2M_{12}\mu_{12} + M_{22}\mu_{22} \quad (8)$$

It is easily seen that, the product of the maximum delay (in seconds) that a bit will incur inside any switch (whether, input-queued, output-queued, or input-output queued) and the rate at which the switch transmits bits at the output port (in bits/sec.) is the bits'-carrying capacity of the switch. This implies that at full capacity, when bits are not being transmitted by the switch, that is, it is on vacation, any incoming bit will be dropped. If therefore, M_{ii} (in seconds) is the maximum delay of a bit in switch i , and μ_{ii} (in bits/sec.) is the maximum rate at which the switch transmits bits, the bits'-carrying capacity of such a switch will be given as $M_{ii}\mu_{ii}$ (bits). Therefore, the capacity of switch 1 is $M_{11}\mu_{11}$ (bits), and $M_{22}\mu_{22}$ (bits) is the capacity of switch 2. However, the capacity of the link connecting the two switches is given by (2), which is, $2M_{12}\mu_{12}$ (the BDP of the link); where, $2M_{12}$ is the maximum round-trip time of the link connecting switches 1 and 2, while the maximum bandwidth of this link is μ_{12} . Hence, the capacity of a 2-switch LAN, which is illustrated in Figure 2, and denoted as CP_2 is given by (8). Similarly, it can be shown that, $\text{Tra}(\boldsymbol{\mu}_2\mathbf{M}_2) = M_{11}\mu_{11} + 2M_{12}\mu_{12} + M_{22}\mu_{22}$; as, it is known generally that, for two square matrices $\mathbf{A}$ and $\mathbf{B}$, $\text{Tra}(\mathbf{AB}) = \text{Tra}(\mathbf{BA})$.

Figure 2.



Likewise, for a 3 switch LAN

$$\mathbf{M}_3 \times \boldsymbol{\mu}_3 = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix} \times \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_{33} \end{bmatrix}$$

$$= \begin{bmatrix} M_{11}\mu_{11} + M_{12}\mu_{21} + M_{13}\mu_{31} & M_{11}\mu_{12} + M_{12}\mu_{22} + M_{13}\mu_{32} & M_{11}\mu_{13} + M_{12}\mu_{23} + M_{13}\mu_{33} \\ M_{21}\mu_{11} + M_{22}\mu_{21} + M_{23}\mu_{31} & M_{21}\mu_{12} + M_{22}\mu_{22} + M_{23}\mu_{32} & M_{21}\mu_{13} + M_{22}\mu_{23} + M_{23}\mu_{33} \\ M_{31}\mu_{11} + M_{32}\mu_{21} + M_{33}\mu_{31} & M_{31}\mu_{12} + M_{32}\mu_{22} + M_{33}\mu_{32} & M_{31}\mu_{13} + M_{32}\mu_{23} + M_{33}\mu_{33} \end{bmatrix}$$

$$\text{Tra}(\mathbf{M}_3\boldsymbol{\mu}_3) = M_{11}\mu_{11} + M_{12}\mu_{21} + M_{13}\mu_{31} + M_{21}\mu_{12} + M_{22}\mu_{22} + M_{23}\mu_{32} + M_{31}\mu_{13} + M_{32}\mu_{23} + M_{33}\mu_{33}$$

By adopting the same reasoning as was explained previously, we have, $M_{12}=M_{21}$, $M_{13}=M_{31}$, $M_{23}=M_{32}$, $\mu_{12}=\mu_{21}$, $\mu_{13}=\mu_{31}$, $\mu_{23}=\mu_{32}$. Hence,

$$\text{Tra}(\mathbf{M}_3\boldsymbol{\mu}_3) = M_{11}\mu_{11} + M_{22}\mu_{22} + M_{33}\mu_{33} + 2M_{12}\mu_{12} + 2M_{13}\mu_{13} + 2M_{23}\mu_{23} \quad (9)$$

Therefore, as was the case above, $M_{11}\mu_{11}$, $M_{22}\mu_{22}$, and $M_{33}\mu_{33}$ are the bits'-carrying capacities of switches, 1, 2 and 3 respectively; while, $2M_{12}\mu_{12}$, $2M_{13}\mu_{13}$, and $2M_{23}\mu_{23}$ are the bits'-carrying capacities of the links or paths between switches 1 and 2, switches 1 and 3, and switches 2 and 3 respectively. Hence, (9), clearly gives the bits'-carrying capacity, CP_3 , of a 3-switch LAN. However, a fundamental property of a network is the bottleneck bandwidth, which sets the upper limit on how quickly the transmitter's data can be delivered to the receiver by the network (Ryousei, Yuetsu, Tomohiro, Motohico, & Fumihiro, 2006; Paxson, 1999); as, according to Lakshman and Upamanyu (1997), the loose definition of BDP is that, it is the product of the RTT (in seconds) for a data connection and the transmission rate (in bits/sec.) of the bottleneck link in the path of that connection. Therefore, assume that each of the links or bits'-pipes on a path or network has bandwidth equal to the smallest link or bits'-pipe's bandwidth, this being the bottleneck link or bits'-pipe: hence, this link's or bits'-pipe's bandwidth is the bottleneck bandwidth on the path or network. It is important in this context to recall here, the minimum capacity postulation with respect to (2). Let it be also assumed that this bottleneck bandwidth upper bounds (5), and that, this bandwidth is denoted by μ . All the bandwidths in (9) can then be replaced with the bottleneck bandwidth; this equation can then be written as:

$$\begin{aligned} CP_3 &= \mu(M_{11} + M_{22} + M_{33} + 2M_{12} + 2M_{13} + 2M_{23}) \\ &= \mu(M_{11} + M_{22} + M_{33} + 2M_{12} + 2M_{13} + 2M_{23}) \end{aligned}$$

In a similar fashion, the capacity of a 2 switch LAN is, $CP_2 = \mu(M_{11} + M_{22} + 2M_{12})$; and the capacity of a 4 switch LAN is

$$CP_4 = \mu(M_{11} + M_{22} + M_{33} + M_{44} + 2M_{12} + 2M_{13} + 2M_{14} + 2M_{23} + 2M_{24} + 2M_{34})$$

It can therefore, be generalized that the bits'-carrying capacity of a switched LAN having n switches, is given by (10). This equation is referred to in this article as the, Eyinagho-Falaki formula for computing any, n -switch LAN's bits'-carrying capacity.

$$CP_n = \mu \left(\sum_{i=1}^n M_{ii} + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n M_{ij} \right) \quad (10)$$

Where, n = number of switches in the LAN; μ = the LAN's bottleneck (that is, smallest) link or bits'-pipe's bandwidth (bits/second); M_{ii} = the maximum delay of any bit in switch i ; M_{ij} = end-to-end maximum delay of a bit in origin-destination path $i \rightarrow j$. Furthermore, assume that the switched LAN is connected to the Internet through switch 1 (recall the illustration of Figure 1); designate this switch as the 'Node for Internet Access', which could be a router or layer 3-switch. In order words, traffic with origins in the Internet, which are destined for hosts in the switched LAN, via other switches (nodes), also in the LAN, must all pass through switch 1. Another name for switch 1 in the context of connecting the switched LAN to the Internet therefore, is, backbone switch. In this context, (10) can be re-written in the form indicated as (11).

$$CP_n = \mu \left(\sum_{i=1}^n M_{ii} + 2 \sum_{j=2}^n M_{1j} + 2 \sum_{i=2}^{n-1} \sum_{j=i+1}^n M_{ij} \right) \quad (11)$$

Very useful physical interpretations can be associated with (10) and (11). The component $2\mu \sum_{i=1}^{n-1} \sum_{j=i+1}^n M_{ij}$ of these equations is largely similar to the widely known definition of a bits'-pipe's capacity, which was previously given in this paper. However, all the origin-destination paths' end-to-end delays of switched LANs, which, traverses the n-interconnected switches have been well accommodated by (10) and (11). The total capacity (in bits) of the paths connecting the: 'Node for Internet Access' to the other nodes of the LAN is given by the second component of (11). Therefore, the minimum capacity (in bits) subscription for Internet access is given by (12). See (2) for the case of a single link or bits'-pipe. Let this minimum capacity be denoted as, MinCap-For-Internet-Access (MCFIA), then,

$$MCFIA = 2\mu \sum_{j=2}^n M_{1j} \quad (12)$$

Approach or approaches for determining the bits'-carrying capacities of switched LANs to the best of the knowledge of the research team for this work, seems not to be available in literature; therefore, literature benchmark(s) for making comparisons are lacking. But to directly prove a proposition (that is, a statement that has a true value – which can, either be TRUE or FALSE), or, an assertion of the form $P \Rightarrow Q$ (Q implies P), involves directly arguing that Q is true, and therefore, P is true (Gross & Yellen, 2006, pp. 681-682; Haggarty, 2002, p. 21). This has been done by the arguments that followed (8) and (9). Moreover, (10), (11), and (12) are proposed, closed-form formulas that are solutions to the problem connected with determining the bits'-carrying capacities of switched LANs, and have no basis for simulation.

One important issue that may pop up after going through this paper, is, the applicability of the formulas that are reported; that is, (10), (11), and (12) to Virtual LANs (VLANs), in the context of computing the bits'-carrying capacities of this class of LANs. The answer to this question is not very difficult. It is a well, known fact that, VLANs are logical structures; which implies that, they are physically wired switched LANs, but, logically configured by software as VLANs. The physical configuration of a switched LAN is therefore, not changed by being logically configured as a VLAN. The logical configuration only creates broadcast domains, that is, it (VLAN) groups stations belonging to one or more physical LANs into broadcast domains. Put differently, when a physically- wired switched LAN is configured as a VLAN, the physical (bits-carrying) capacity does not change: this is because, the number of physical hosts or stations, and physical switches, does not change, so also, is the physical wiring, which is not changed. Therefore, the practical utility of the formulas is not changed by the fact that, a switched LAN is configured as a VLAN or not.

An Example to Illustrate the Computation of the Bits'-Carrying Capacity of a Switched LAN

Covenant University, Nigeria, College of Science and Technology (CST) building, switched LAN is used to illustrate the applicability of (10) and (12). It is shown in Figure 3. The LAN has eight switches and

one router. One of the switches (CISCO 3550) is serving as a backbone switch to six other CISCO 2950 switches. A seventh CISCO 2950 switch connects the CISCO 2600 router and the CISCO 3550 switch. It looks like the seventh switch – CISCO 2950 is redundant. There is a DVB (Digital Video Broadcasting) device, used for digital video signals' reception. The CISCO 2600 (which is a router) connects to the VSAT (Very Small Aperture Terminal) through a Radyme Comstream modem. The end-users' hosts or stations that are located in the CST building are interconnected using the CISCO 2950 switches.

Figure 3.

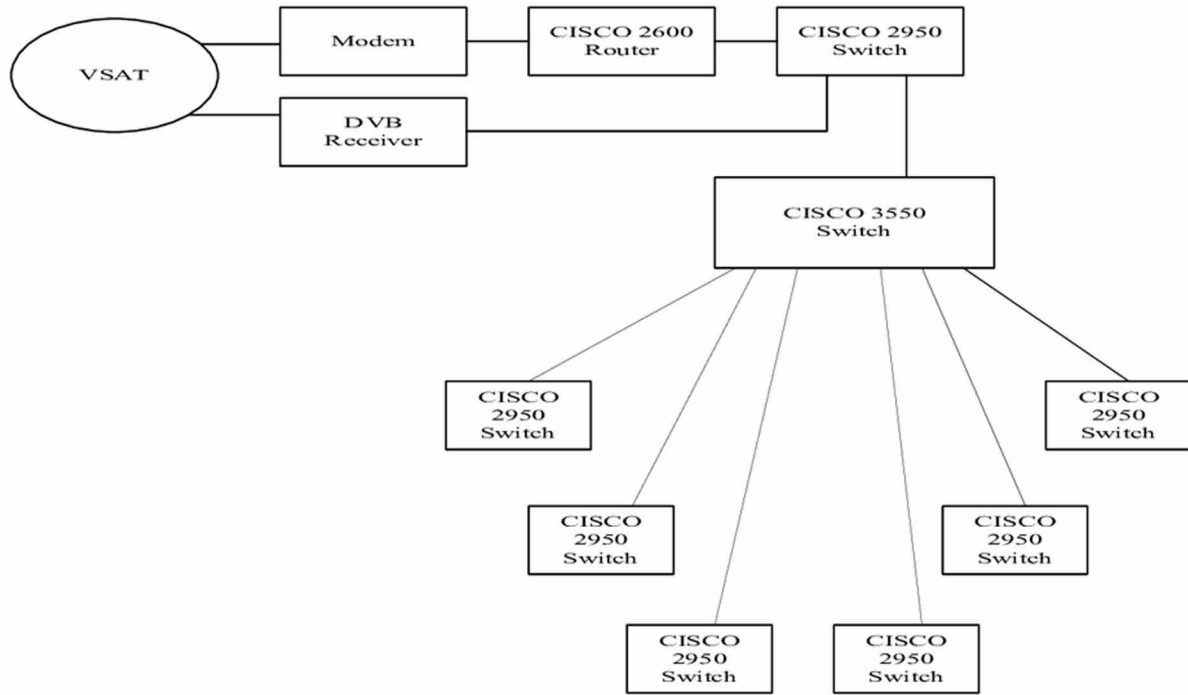


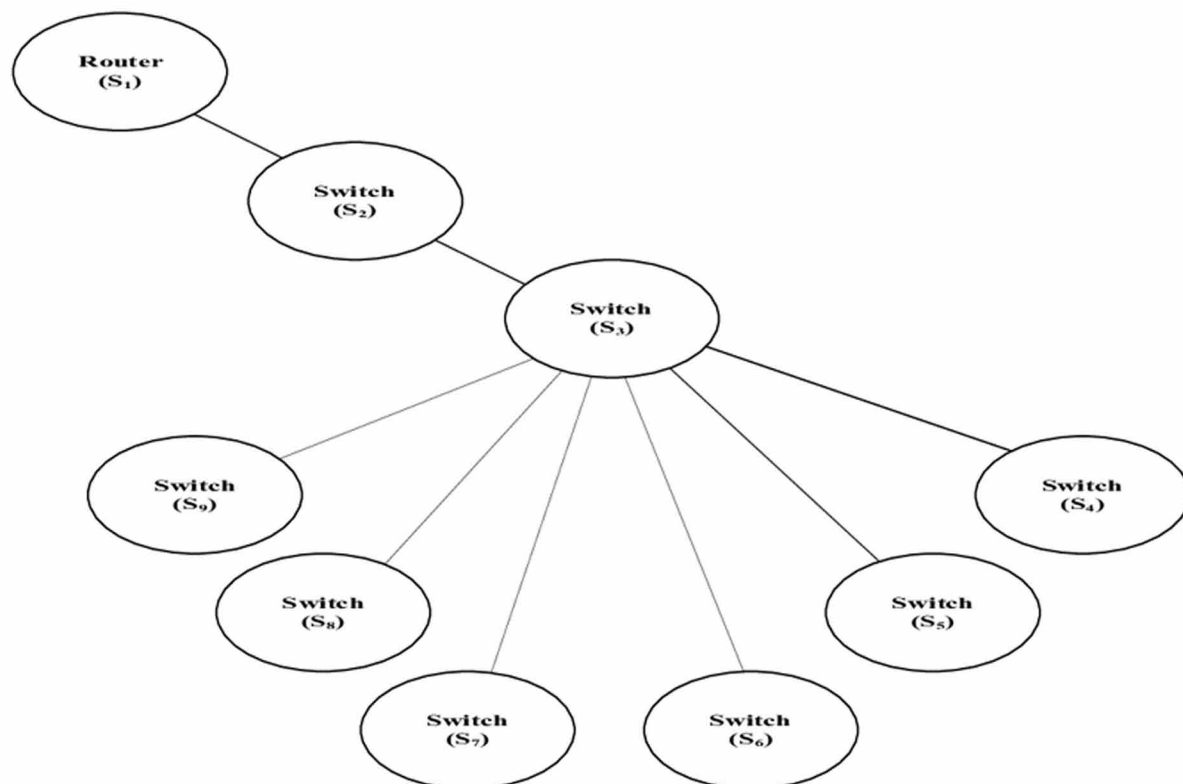
Figure 3 was drawn with the attached hosts not included in the drawing. This is in consonance with the fact that, as was shown by Eyinagho, Falaki, and Atayero (2011), the number of end-to-end delays of a switched LAN only depends on the number of switches in the LAN. It should also be noted that, a router is a switching device, which, in the context of this switched LAN, is taken as one of the switches. Figure 4 shows Figure 3 redrawn, with the DVB receiver, the modem and the VSAT terminal omitted.

Take, $D_{2950\max}$ = the maximum delay of an Ethernet packet in the CISCO 2950 switch, $D_{3550\max}$ = the maximum delay of an Ethernet packet in the CISCO 3550 switch, and $D_{2600\max}$ = the maximum delay of an Ethernet packet in the CISCO 2600 router; it was shown in (Eyinagho, Falaki, & Atayero, 2012, pp.169-171) that, $D_{2950\max} = 41.91945\text{ms} \cong_{42\text{ms}}$, $D_{3550\max} = 41.92\text{ms} \cong 42\text{ms}$, $D_{2600\max} = 41.89\text{ms} \cong 42\text{ms}$. As there are nine (9) switches in the LAN, (10) now becomes,

$$CP_9 = \mu \times \left(\sum_{i=1}^9 M_{ii} + 2 \sum_{i=1}^{9-1} \sum_{j=i+1}^9 M_{ij} \right) = \mu \times ([M_{11} + M_{22} + M_{33} + M_{44} + M_{55} + M_{66} + M_{77} + M_{88} + M_{99}] +$$

[illegible]

Figure 4.



Minimum Bits Capacity Subscription for Internet Access' Computation

Bits' capacities are usually offered by ISPs to their customers in units such as Megabits (Mb), Gigabits (Gb). The computation of this capacity can be performed with Eq. (12). For this switched LAN, $n=9$, therefore,

$$\text{MCFIA (MinCap-For-Inter-Access)} = 2\mu \sum_{j=2}^n M_{1j} = 2 \times \mu \times \sum_{j=2}^9 M_{1j}$$

$$= 2 \times \mu \times [M_{12} + M_{13} + M_{14} + M_{15} + M_{16} + M_{17} + M_{18} + M_{19}] = 2 \times 10 \times 10^6 \text{ bps} \times [84 + 126 + 168 + 168 + 168 + 168 + 168 + 168] \text{ ms} \\ = 20 \times 10^6 \text{ bps} \times 1218 \text{ ms} = 20 \times 10^6 \text{ bps} \times 1.218 \text{ secs} = 24,360,000 \text{ bits} = 24.36 \text{ Mbits.}$$

This is the minimum (and hence, optimum) upload and download, bits' capacity which, for seamless Internet access, is recommended for purchase from an ISP in order for the users whose hosts are connected to this switched LAN to have good QoE (Quality of Experience), during upload and download of information to and from, the Internet. Purchasing above this bits' capacity is obviously a waste of financial resources, except of course, there is an increase in the number of hosts (as a result of increase in the user-base), necessitating the installation of more switches. Purchasing below this capacity will lead to increases in delay, during Internet upload and download sessions, resulting in poor users' QoE.

FUTURE RESEARCH DIRECTIONS

It is proposed here that, utilizing the formulas reported in this article for computing the bits'-carrying capacities of switched LANs, and, the capacities to be subscribed for, when switched LANs are to be connected to the networks of ISPs for the purposes of Internet accesses, presents economic and QoE advantages, over, rule-of-thumb approaches. Further research efforts will now be geared towards carrying out studies to test this proposition.

CONCLUSION

The development of empirical formulas for computing the bits'-carrying capacities of switched LANs, and the bits' capacities, in, for example, Mega-Bytes, Giga-Bytes, to be subscribed for, if there is the need to connect such LANs to the Internet has been reported in this article. Applying any of the formulas is quite simple, requiring mainly, the computation of all the maximum end-to-end delays of a switched LAN. This can be done using, methodologies and formulas which were developed and reported in previous works, as cited in this article. The formulas are in the main, quite novel, as, to the best of the knowledge of the researchers involved in this work, no such practically, utilizable formulas seems to exist in literature. Based on this fact, the researchers resorted to proving, using simple inductive principles, but with quite understandable, physical interpretations, the practical and advantageous utility of the reported formulas.

REFERENCES

- CISCO CCNA Exploration Companion Guide. (n.d.). *LAN Switching and Wireless*. Available: www.ciscopress.com
- Comer, D. (2004). *Computer Networks and Intranets with Internet Applications*. Pearson Prentice Hall.
- Eyinagho, M. O. (2012). Determination of the End-To-End Delays of Switched Local Area Networks. *International Journal of Computer and ICT Research*, 6, 24–30.
- Eyinagho, M. O., Falaki, S., & Atayero, A. (2011). Revisiting the notion of Origin-Destination Traffic Matrix of the hosts that are attached to a Switched Local Area Network. *International Journal of Distributed and Parallel Systems*, 2(6), 227–235. doi:10.5121/ijdps.2011.2620
- Eyinagho, M. O., Falaki, S., & Atayero, A. (2012). *Determination of End-To-End Delays of Switched Local Area Networks*. LAP LAMBERT Academic Publishing GmbH and Co.
- Forouzan, B. (2008). *Data Communications and Networking*. McGraw-Hill Limited.
- Forum, M. E. (n.d.). *Understanding Carrier Ethernet Throughput*. Available: www.metroethernetforum.org
- Gerd, K. (1989). *Local Area Networks*. McGraw-Hill Book Co.
- Gross, L., & Yellen, J. (2006). *Graph Theory and Its Applications*. Chapman and Hall.
- Haggarty, R. (2002). *Discrete Mathematics for Computing*. Pearson Education Ltd.
- Kai, C., Yuan, X., Samarth, H. S., & Klara, N. (2004). Understanding Bandwidth-Delay Product in Mobile Ad Hoc Network. *Computer Communications*, 27(10), 923–934. doi:10.1016/j.comcom.2004.01.020
- Lakshman, T. V. & Upamanyu, M. (1997). The Performance of TCP/IP for Networks with High Bandwidth-Delay Products and Random Loss. *IEEE/ACN Transactions on Networking*, 5(3), 336–350.
- Paxson, V. (1999). End-to-End Internet Packet Dynamics. *IEEE/ACM Transactions on Networking*, 7(3), 277–292. doi:10.1109/90.779192
- Rangaswamy, S., & Krishnareddy, V. (2015). An Efficient Traffic Regulation Mechanism for Distributed Networks. *EURASIP Journal on Wireless Communications and Networking*, 2015(1), 154. doi:10.1186/13638-015-0386-2
- Rathi, S., Malik, N., Chahal, N., & Malik, S. (2014). Throughput for TDD and FDD 4G LTE Systems. *International Journal of Innovative Technology and Exploring Engineering*, 3(12).
- Ryousei, T., Yuetsu, K., Tomohiro, K., Motohico, M., & Fumihiro, O. (2006). Real Time Burstiness Measurement. *Proc. PFLDnet Conf.*, 109–115.
- Summers, J., Barford, P., Greenberg, A., & Willinger, W. (2008). An SLA Perspective on the Router Buffer Sizing Problem. *Performance Evaluation Review*, 35(4), 40–51. doi:10.1145/1364644.1364645
- Szymanski, T. H. (2016). *Securing the Industrial-Tactile Internet of Things with Deterministic Silicon Photonic Switches*. *IEEE Access Magazine*. doi:10.1109/ACCESS.2016.2613512
- Tanenbaum, A. (2006). *Computer Networks*. New Delhi, India: Prentice-Hall of Indian Private Limited.

Zubair, S., Ahmad, S., Faraz Ahmad, B. L., & Ali, A. (2014). Novel Approach for Bandwidth Optimization in Grid-Computing Infrastructure. *International Conference on Computer Science and Network Security (CSNS 2014)*.

ADDITIONAL READING

Alberto, L., & Widjaja, I. (2004). *Communications Networks: Fundamental Concepts and Key Architectures*. McGraw Hill.

Alexander, S., Eitan, Z., & Ori, R. (2014). The Buffer Size Vs Link Bandwidth Trade-off in Lossless Networks. *IEEE Hot Interconnects: Symposium on High Performance Internetworks*, Mountain View, CA.

Anurag, K., Manjunath, D., & Kuri, J. (2004). *Communication Networking: An Analytical Approach*. San Fransisco, USA. Morgan Kaufmann Publishers.

Bertsekas, D., & Gallager, R. (1992). *Data Networks*. Prentice-Hall.

Ma, Y., & Jia, Z. (2017). Evolution and Trends of Broadband Access Technologies and Fiber-Wireless Systems. In M. Tornatore, G.-K. Chang, & G. Ellinas (Eds.), *Next-Generation Communication Networks: Systems, Architecture and Management* (pp. 45–75). Springer International Publishing. doi:10.1007/978-3-319-42822-2_2

Ming-Yang, X., Rong, L., & Huimin, C. (2004). Predicting Internet End-to-End Delay: An Overview. *Proceedings of the IEEE 36th South Eastern Symposium on Information Systems Theory*, 210 – 214.

Rainer, G., & Wischik, D. (2005). Buffer Sizes for Large Multiplexers: TCP Queuing Theory and Instability Analysis. In *EuroNGI*, Rome, Italy.

Wenfeng, X., Yonggang, W., Chuan, H.-F., Dusit, N., & Haiyong, X. (2015). A Survey on Software-Defined Networking. *IEEE Communications Surveys and Tutorials*, 17(1), 27–51. doi:10.1109/COMST.2014.2330903

KEY TERMS AND DEFINITIONS

Bandwidth Delay Product (BDP): This is the product of the Round-Trip Time (RTT) of a link and its bits'-transmission capacity.

Bits'-Carrying Capacity: The maximum number of bits that a network can accommodate at any point in time, without, an abnormal increase in response time, and, decrease in throughput.

End-to-End Delay: This is the time taken for a bit or packet to traverse from a transmitter to a receiver.

Network Performance Management System: An application program that is used to, manage the performance of a communication network.

Node for Internet Access: The router or layer-3 switch that connects an organization's switched Local Area Network to the world-wide Internet.

Round-Trip Time (RTT): This is the time taken by a bit to traverse from a transmitter to a receiver, and back to the transmitter.

Switched Local Area Network: A Local Area Network with the users' hosts interconnected by switching devices, like, switches and routers.

Design and Analysis of AWG and FDL-Based Optical Switch in Data Centre Network

Vaibhav Shukla

Tech Mahindra, Mumbai, India

Dilip Kumar Choubey

 <https://orcid.org/0000-0002-1233-7159>

Vellore Institute of Technology, Vellore, India

INTRODUCTION

The fast growth in modern computing applications such as video on demand, multimedia application use of the internet of things generates thousands of discrete data in every minute. This rapid growth of data generation, generates a heavy workload on the data center servers. Major content provider companies such as yahoo, google, Microsoft, amazon, etc, has installed mega data center systems which house hundreds of thousands of servers in large scale layouts. (Habib, 2012; Koley, 2010; Katz, 2009). In the network as the demand of network services increases the requirement of high bandwidth also increases. The electronic devices do not fulfil this demand of higher bandwidth because of speed constraints (Sato and Hasegawa, 2009). So, the use of optical technology plays a significant role. Now days to meet this increasing demand in communication network the optical network technologies are explored by researchers. The one of the promising techniques is an optical packet switched network. In an optical packet switching system all switching functions are performed optically. The optical network system consists of core and client networks. In optical switching system, this is difficult to implement the control and processing units in optical domains and this is not technical feasible. So, in most of the optical switching system, the buffering and switching are performed optically while all the control functions are performed in electronic domain.

The main objective of this article is to discuss a wavelength division multiplexed, optical switch. The physical layer and network layer analysis of switch is presented in the article. The switch is placed in to the network and the overall performance of switch is measured under different conditions when it is placed in to the network.

BACKGROUND

In this section a brief background of optical network is presented the various devices which are used in to the core of optical switches is also discussed in this section.

DOI: 10.4018/978-1-7998-3479-3.ch068

Advantages of Optical Communication

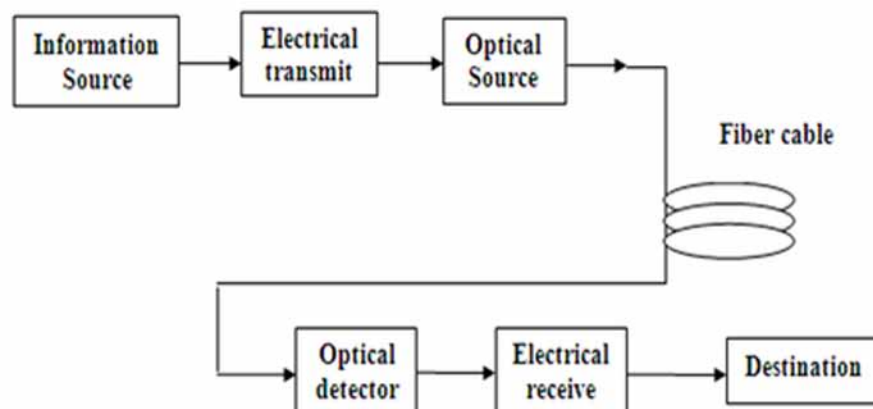
The optical fiber technology is based on the propagation of the optical pulses in optical fiber. Hence, the system offers advantages like large optical bandwidth (theoretically infinite), shock proof immune to interference etc. The advantages of the optical fiber communication can be summarized as (Agrawal, 2012):

1. The optical fiber cable offers a very large bandwidth, it is very light weight, and diameter is very small, hence laying of cable is not a daunting task.
2. The optical fiber cable is non-conductor; hence systems are shock free.
3. The optical fibre systems are immune to Interference and Crosstalk
4. The transmission is very secure and fiber lay down under ground and tapping of the signal is not possible. Thus, it offers very high security of signals.
5. In the region of interest the transmission loss is negligible.
6. The optical fiber communication systems are rugged and very flexible.
7. The optical fiber communication systems are very reliable and maintenance is easy.

GENERIC OPTICAL FIBER COMMUNICATION SYSTEM

In 1960, during his research for US Air force, RAND corporation scientist Paul Baran was the first one to introduce the concept of packet switching in 1960 (Agrawal, 2012). In packet switching, the information is transferred into small size packets. Each packet contains small chunks of data which is to be transmitted between various nodes into network. Inside the network, at each node the packets are either buffered or it can be directly transmitted towards the output port which results into variable queuing delay among packets. This feature makes a difference between circuit and packet switching as in circuit switching a dedicated path is established in advance and continue till communication takes place. For optimizing available network channel capacity, packet switching technologies are used (Agrawal, 2012). By using packet switching, the transmission latency can be minimized (i.e. the time taken by data to pass across the network), and the robustness of network can be increases (Koch, 2012).

Figure 1. The generic optical fiber communication system.



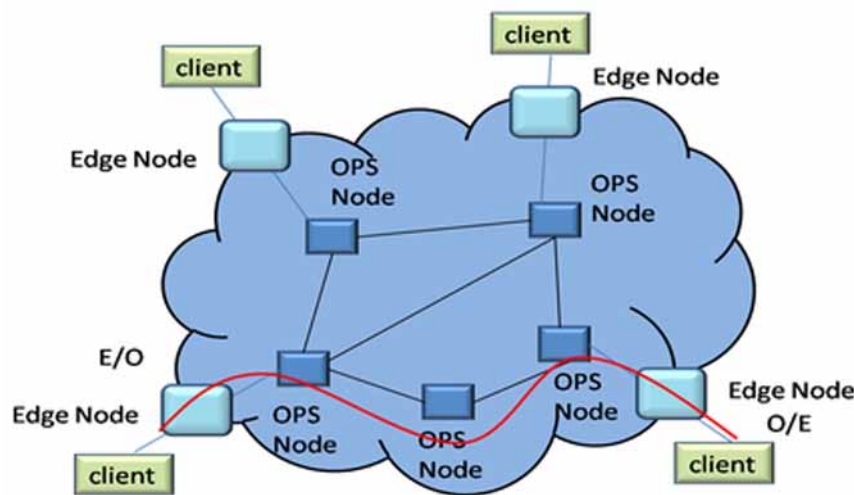
Optical packet switching (OPS) is a switching technology with fine granularity. OPS technology transmits data in optical domain without O/E and E/O conversion (Koch, 2012). The key issue involved in the OPS is the design of the switch/router which can perform switching of the data at a very high data rate. In this article, design and analysis of AWG based optical packet switch is detailed with its advantages and limitations. The performance evaluation of the switch is done at both physical and network layer.

The generic optical fiber communication system is shown in figure.1. The information is generated in the form of an electrical signal, generally derived from a real message signal which may or may not be electrical (e.g. sound), to a transmitter comprising electrical and electronic components which converts the signal into an appropriate signal format for propagation over the transmission medium. The optical carrier generally modulated with electrical and optical modulator. On the other hand, at the receiver the optical signal is again converted back into the electrical signals and directed towards destination (Habib et al., 2012). The above depicted figure 1 only shows a single fiber link, between the source and destination. However, in networks thousands of such links exist. A simple optical network layout is discussed in next section.

OPTICAL NETWORKS

The figure 2 shows the generic layout of network. The network structure contains the client and the core networks. In between client and core network the edge routers are used as an interface. These edge router lies on the periphery of the network cloud.

Figure 2. The generic layout of the OPS system.



In network currently both core and edge routers are electronic in nature. The major drawback of electronic switches/routers is that it has limited speed (Tucker, 2008). Therefore, in high data rate transmission the use of electronic routers is not feasible. This inadequacy of the electronic routers gives rise to birth of optical networks. The major issue involved in the design of optical networking is the design of the switch routers which can perform efficiently in the high data rate network. These can be classified as

‘all’ - optical or photonic switches (Tucker, 2008). In first category i.e. in all optical mode, both the data propagation and processing are performed in optical domain. In current industry due to the unavailability of optical RAM all optical switches are not technically feasible (Tucker, 2008). In the other mode the data transmission is performed in optical domain while the control operations are performed in electronic domain. By using photonic packet switching in network, high speed transmission, format transparency efficiency and flexibility in the configuration is achieved. In next generation telecommunication system, the edge routers work in electronic domain and the core routers will be implemented optically. This hybrid approach allows the efficient utilization of the mature electronic technology and huge bandwidth of the fiber in optical domain. In this approach, it is assumed that edge routers are capable of packet aggregation; means they can handle variable length packets arriving from the various client networks. Finally, edge routers push the fixed length packets in the core routers networks.

TRANSMISSION TECHNOLOGIES

In optical transmission two switching technologies, optical packet switching and optical burst switching are heavily investigated in past (Yoo, 2006; Xuy, Perros and Rouskas, 2001; Mujherjee, nd; Stern and Bala 1999; Senior, 2004). In optical burst switching system data units are transferred in to the form of burst of packets while in optical packet switching system the data transmission is performed in the form of small size of packets of same or variable length. The detailed discussion of optical packet switch technologies is as given below:

Optical Packet Switching

In 1960, during his research for US Air force, RAND corporation scientist Paul Baran was the first one to introduce the concept of packet switching in 1960. In packet switching, the information is transferred into small size packets. Each packet contains small chunks of data which is to be transmitted between various nodes into network. Inside the network, at each node the packets are either buffered or it can be transmitted towards the output port directly which results into variable queuing delay among packets. This feature makes a difference between circuit and packet switching as in circuit switching a dedicated path is formed in advance and continue till communication occurs. For optimizing available network channel capacity, packet switching technologies are used. By using packet switching, the transmission latency can be minimized, and the robustness of network can be increases.

Optical packet switching (OPS) is a switching technology with fine granularity. OPS technology transits data in optical domain without O/E and E/O conversion (Yoo, 2006; Xu, Perros and Rouskas, 2001). The key issue involved in the OPS is the design of the switch/router which can perform switching of the data at a very high data rate. In this thesis, design and analysis of AWG based optical packet switch is detailed with its advantages and limitations. The performance evaluation of the switch is done at both physical and network layer.

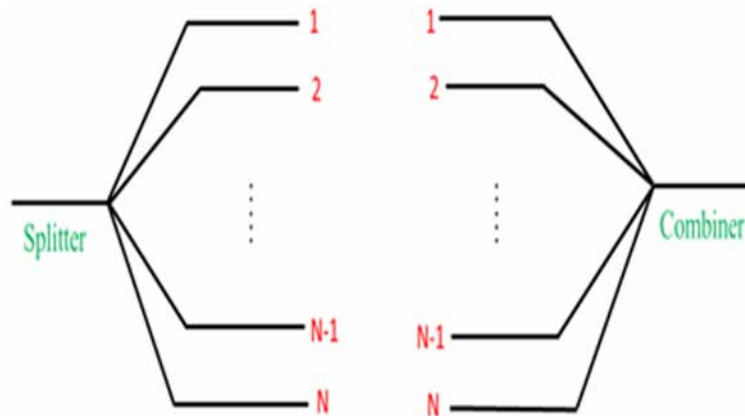
OPTICAL SWITCH DESIGN AND COMPONENTS

In the design of the optical switch or networks various optical components are needed. Over, here only those components are discussed which are used in the work presented in the article.

Splitters and Combiners

The generic layout of splitter and combiner is shown in figure 3. The splitter is a device which is used to divide the power of incoming signal in equal parts whereas on the other hand, the combiner is used for the opposite. The splitter and combiners are wavelength insensitive devices.

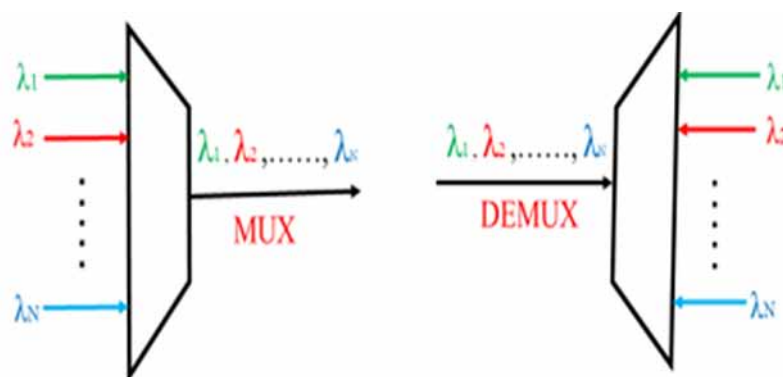
Figure 3. Schematic of Splitter and Combiner.



Multiplexer and De-multiplexer

For combining and separating light signals at different wavelength, MUX and DEMUX are used in optical devices. These two devices are used as wavelength sensitive devices. In general, WDM MUX/DEMU are based on arrayed waveguide grating or interferometric principle. The figure 4 shows the design of Mux and Dmux.

Figure 4. Schematic of Multiplexer and Demultiplexer.



Arrayed Waveguide Gratings (AWG)

The AWG router is used to map an input waveguide to an array of output waveguides in such a way that the different wavelength signals present in the input waveguide are mapped on the different output waveguides (Kato, et al., 2000). The AWG works on the interferometer phenomenon. In wavelength division multiplexing optical systems, they are used as optical multiplexers/ demultiplexers. The figure 5 shows the design of AWG router. The arrayed waveguide grating router follows a definite routing pattern of different wavelengths. The figure 6 shows a routing pattern of 4×4 AWG.

Figure 5. Representation of Arrayed Waveguide Grating.

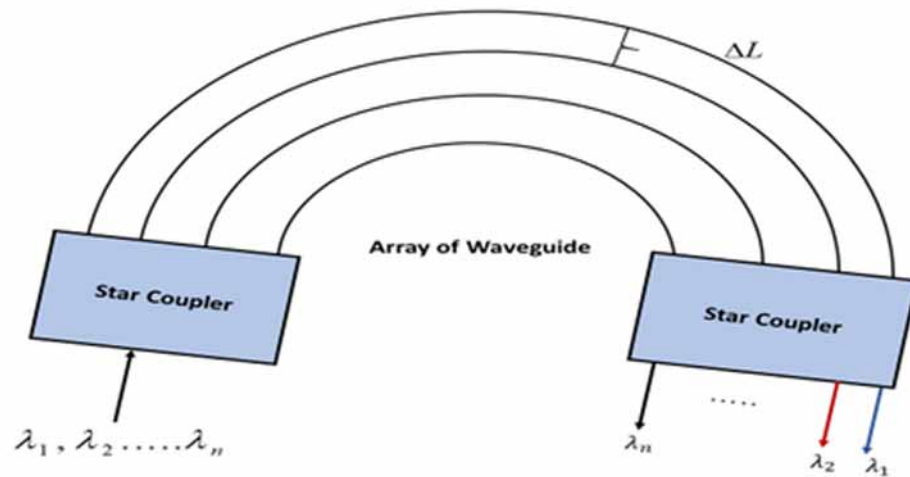


Figure 6. Representation of routing pattern of 4×4 AWG router.

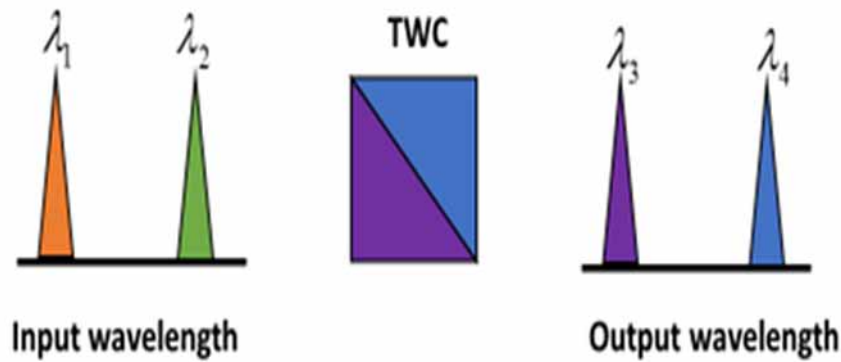
Routing Pattern		Outputs			
		1	2	3	4
Inputs	1	λ_1	λ_2	λ_3	λ_4
	2	λ_2	λ_3	λ_4	λ_1
	3	λ_3	λ_4	λ_1	λ_2
	4	λ_4	λ_1	λ_2	λ_3

Tunable Wavelength Converter (TWC)

The invention provides a tunable wavelength converter that converts an optical signal from one wavelength to another set of wavelengths as shown in figure 7. In general wavelength converters are expected to have the following features:

1. Large wavelength conversion capacity
2. Fast wavelength tuning time
3. Polarization independent
4. High Signal to noise ratio.
5. Transparency to bit rates and signal formats.

Figure 7. Design of Tunable Wavelength Converter



Optical Amplifier

Optical Amplifier is a device which is used to amplify the optical signal directly without converting it into an electrical signal. In general, two types of optical amplifiers are used commonly.

Semiconductor Optical Amplifier

Semiconductor Optical Amplifier (SOA) is a type of amplifier which is used to amplify the optical signals into optical domain. The gain and noise factors in the case of SOA are represented as:

The gain and noise factor in semiconductor optical amplifier are modelled as:

$$G = G_0 \exp \left(- (G - 1) \frac{P_{in}}{P_{sat}} \right)$$

$$S_{sp} = \eta_{sp} h\nu (G - 1) \text{ Watts/Hz}$$

In this equation, G represents SOA gain, G_0 is used to represent unsaturated gain, P_{in} represents the input power enters in to SOA, P_{sat} shows saturation power level, S_{sp} represents amplifier spontaneous emission (ASE) noise power spectral density, n_{sp} represents spontaneous emission factor, h is Plank's constant and ν represents optical frequency.

EDFA (Erbium Doped Fiber Amplifier)

In the case of long range optical fiber communication, EDFA is the most important amplifier. By using EDFA, we can efficiently amplify the light signal in 1.5 μm wavelength region, where telecom fiber has the minimum loss (Ramaswami, Sivarajan and Sasaki, 2009). As like in SOA, the EDFA amplify the optical signals into the optical domain so that there is no need of expensive O/E and E/O convertors. The figure 8 shows the generic layout of EDFA. The table 1 shows the difference between both amplifiers' SOA and EDFA.

A piece of fiber is used in EDFA, whose core is doped with Erbium ions.

Advantages of EDFA

1. High Gain
2. Low Noise
3. Low temperature sensitive
4. Excellent Coupling

Disadvantages of EDFA

1. Pump laser is required
2. Integration with different components is difficult
3. Dynamic control is required

Figure 8. Schematic diagram of EDFA.

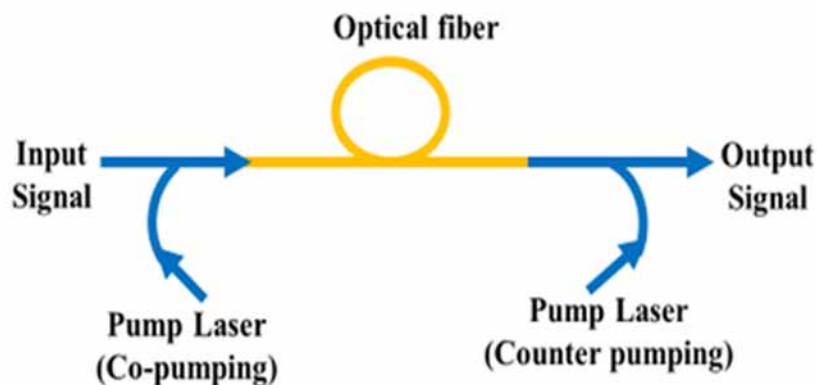


Table 1. Difference between EDFA and SOA.

Property	EDFA	SOA
Gain (dB)	>40	>30
Wavelength	1530-1560	1280-1650
Noise Figure	5	8
Size	Rack Mounted	Compact
Cost Factor	Medium	Low
Switchable	No	Yes

SOME PROMINENT AWG BASED OPTICAL SWITCH DESIGNS

In this section some recently published AWG based optical packet switch architectures are discussed in brief.

In 2013 the H. Rastegarfar (2012) presents an AWG based optical packet switch architecture the switch is a recirculating loop buffer-based switch. In this design when packet arrives at the input of switch the TWC present at the input converts the wavelengths of incoming packets as per the routing pattern of AWG. Here if the packet feels contention then it is forwarded towards the recirculating loop buffer where packets can be stored. After one circulation if the packet still feels contention then it is again forwarded towards the direction of loop buffer other wise it is forwarded towards the output port of switch.

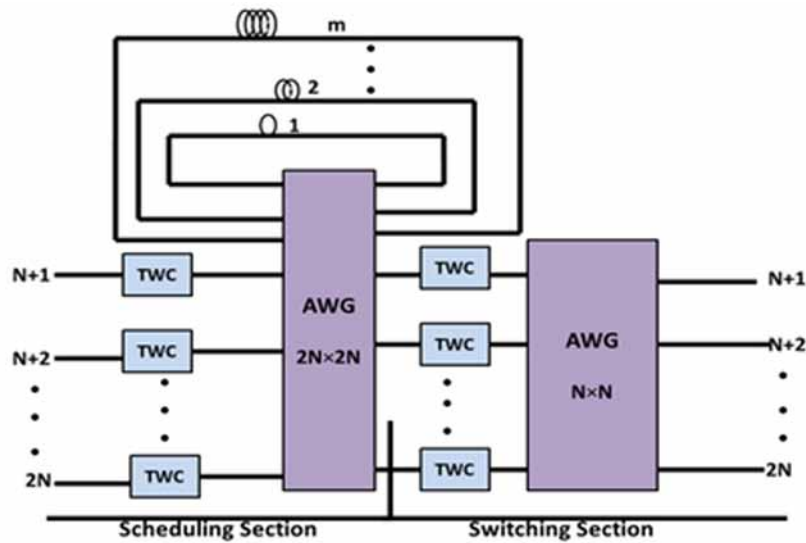
In 2013 Y. Yin et.al. (2012) proposed a buffer less NACK based optical switch design in which due to absence of effective buffering mechanism a large number of packets will be re-transmitted. In 2017 P. Bhattacharya et.al (2017) proposed a design having a $N \times N$ AWG router and N input / output lines used as the core of switch. The data receives at the input lines are in the form of packets. In design each input is connected with a TWC which tunes the wavelength of incoming packets as per the routing pattern of AWG.

In 2018 A. Singh et.al. (2018) proposed a switch design. This switch uses a hybrid optical buffer that merge the advantages of both fast-electronic buffering and fiber optical buffering. In this design numerous sets of fiber loops and electronic buffering modules are joined and develop the hybrid optical buffer. Initially the packets are stored in to the optical buffer, in a temporarily basis. In the event that packet cannot be sent out after a particular interval, the conversion of packet will be done and it will be stored in electronic memory.

ARCHITECTURE DESCRIPTION AND PROPOSED MODIFICATIONS

In figure 9 an AWG based switch is presented in which fiber delay lines are used as an optical storage (Shukla and Srivastava, 2015). The whole switch is divided in to two sections scheduling and switching sections. The control operations in the switch are performed through electronic controller and data transmission is performed optically. In the scheduling section of switch, at input a set of TWC is present which is used to convert the wavelength of incoming packets as per the routing pattern of first AWG. The first N ports of scheduling AWG are linked with fiber delay lines these delay lines are used to store the contended packets. The other $N+1$ to $2N$ ports are worked as actual input-output port of the architecture. In switching section, a set of TWC is present just after first AWG these TWC are used to convert

Figure 9. Schematic of AWG based switch design (D1) (Singh & Srivastava, 2018).



the wavelength of incoming packets as per the routing pattern of switching section AWG and at last the switching section AWG forwarded the packets towards its appropriate destination port. The total number of buffer modules present in the switch are “ m ” and it is to note that the maximum value of “ m ” is N .

Problems in Switch D1

Some of the major Problems associated with the architecture D1 are as follows:

- The buffered packet has the poor quality of signals in comparison to the direct transferred packets.
- As the length of each buffer unit is different so buffer packet also suffers different loss.
- The optimum receiver design is not possible, as different packet suffers different loss.

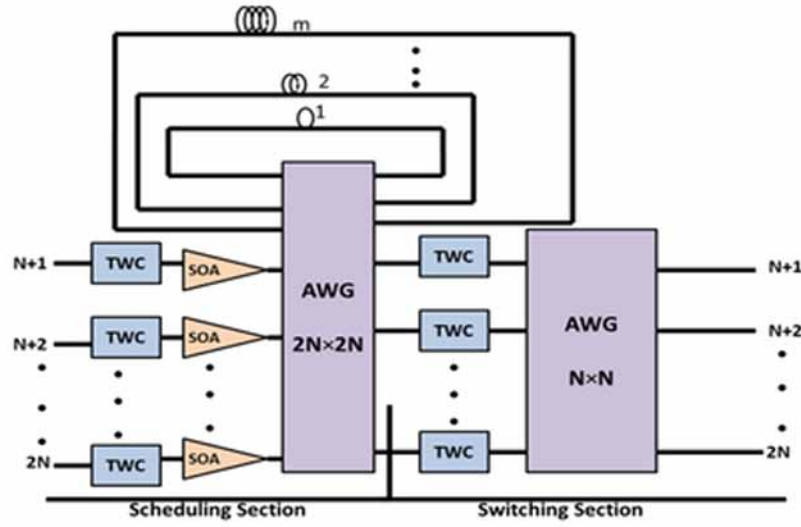
Due to the problems associated with the architecture D1, there is a need to modify the switch architecture. Hence, modification in the design of switch is proposed and the performance in terms of loss, power and cost are presented in to the paper.

Architecture Modifications

In modification an SOA is placed at each input branch of scheduling section AWG as shown in figure 10. Here the loss of the loop as well as directly transferred packets are fully compensated by the corresponding amplifier available at each input of switch. The advantage of this modification is that, SOA is a noisy device and some of noise generated by SOA are reduced by other components.

The figure 10 shows a modified architecture this modification makes sure that each of the packet have same optical power once they received at the output of switch. Otherwise some of the packets have better signal quality in comparison to others, and it may happen that some of the packets may not have sufficient power that can be detected at the output of switch.

Figure 10. Schematic of AWG based switch design (D2) (Srivastava & Singh, 2010).



ANALYSIS OF SWITCH D2

In this section the detailed loss, power analysis of switch D2 (figure 4) is presented.

Case1: Loss analysis of switch D2 when packet passes through the buffer

The total loss of switch D2 when packet passes through the buffer is given by eq. (1):

$$A_T = A_{TWC} A_{SOA} A_{FDL} A_{AWG}^{2N \times 2N} A_{TWC} A_{AWG}^{N \times N} A_{AWG}^{2N \times 2N} \quad (1)$$

Case2: Loss analysis of switch D2 when packet directly passes through the output unit

The total loss of switch D2 when packet directly passes through the output unit is given by eq. (2):

$$A_T = A_{TWC} A_{SOA} A_{AWG}^{2N \times 2N} A_{TWC} A_{AWG}^{2N \times 2N} \quad (2)$$

Power Analysis of Switch D2

Power entering in buffer module for bit b is given by eq. (3)

$$P_s = bP_{in} \quad b \in [0,1] \quad (3)$$

Power at the output of the switch D2 is given by eq. (4)

$$P_{out}^{D2} = P_{in} + n_{sp}(G-1)h\nu B_0 A_{TWC} A_{AWG}^{N \times N} A_{TWC} \quad (4)$$

Noise Analysis of Switch D2

At receiver, different noise components for the bit “b” are presented through the set of eq. (5):

Shot noise

$$\sigma_s^2 = 2qRPB_e$$

ASE-ASE beat noise

$$\sigma_{sp-sp}^2 = 2R^2P_{sp}(2B_o - B_e)\frac{B_e}{B_0^2}$$

Sig-ASE beat noise

$$\sigma_{sig-sp}^2 = 4R^2P\frac{P_{sp}B_e}{B_0}$$

Shot-ASE beat noise

$$\sigma_{s-sp}^2 = 2qRP_{sp}B_e$$

Thermal noise

$$\sigma_{th}^2 = \frac{4K_BTB_e}{R_L} \quad (5)$$

The various noise components will be given by eq. (5), the expression for P and P_{sp} will be given by eq. 6-7

$$P = bP_{in} \quad (6)$$

$$P_{sp} = P_{in} + n_{sp}(G-1)h\nu B_0 A_{TWC} A_{AWG}^{N \times N} A_{TWC} \quad (7)$$

The total noise variance for bit b is given by eq. (8)

$$\sigma^2(b) = \sigma_s^2 + \sigma_{sp-sp}^2 + \sigma_{sp-sig}^2 + \sigma_{s-sp}^2 + \sigma_{th}^2 \quad (8)$$

The BER can be obtained from eq. (9-10)

$$BER = Q\left(\frac{I(1) - I(0)}{\sigma(1) - \sigma(0)}\right) \quad (9)$$

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-\frac{z^2}{2}} dz \quad (10)$$

CALCULATIONS

The table 1 shows the different components and their values used in the calculation.

Table 2. Used parameters and their values (Srivastava, Singh, & Singh, 2008; Shukla, Sonkar, & Srivastava, 2019).

Parameter	Value
Switch Size (N)	4,8
Population inversion factor (n_{sp})	1.2
Amplifier gain (G)	20 dB
Speed of light (c)	$3 \times 10^8 \text{ m / s}$
Refractive index of fiber (n)	1.55
Responsively (R)	1.28 A/W
Electronic charge (e)	$1.6 \times 10^{-19} \text{ C}$
Electrical bandwidth (B_e)	20GHz
Optical bandwidth (B_0)	40GHz
TWC insertion loss (L_{TWC})	2.0 dB
Loss of Scheduling and Switching AWG $L_{AWG}^{2N \times 2N}$	3.0 dB
Loss of the fiber loop (L_{FDL})	0.2 dB/km
Loss of SOA (L_{SOA})	1 dB

Loss of the Architecture D2

Inserting the values of the various loss parameters we get the total loss of switch as shown in eq. (11-12),

When packet passes through buffer

$$A_b^T = 15.6 \text{ dB} \quad (11)$$

For straight through packets

$$A_r^S (\text{dB}) = 11.0 \text{ dB} \quad (12)$$

Here eq. (11) represents the loss through the buffer and eq. (12) straight path loss in switch D4. It is obvious that the loss in both the architecture is same due to the same number of components. However, in this switch each packet is passes through SOA so in each case when packet passes through buffer or straight through loss are compensated through SOA as shown in eq. (13).

$$A_r G_{\text{SOA}} = 1 \quad (13)$$

The total loss of the switch D2 is given by eq. (14)

$$A_r = 0 \text{ dB} \quad (14)$$

However, the noise value of each switch will be different as the position of SOA in both the architecture is different.

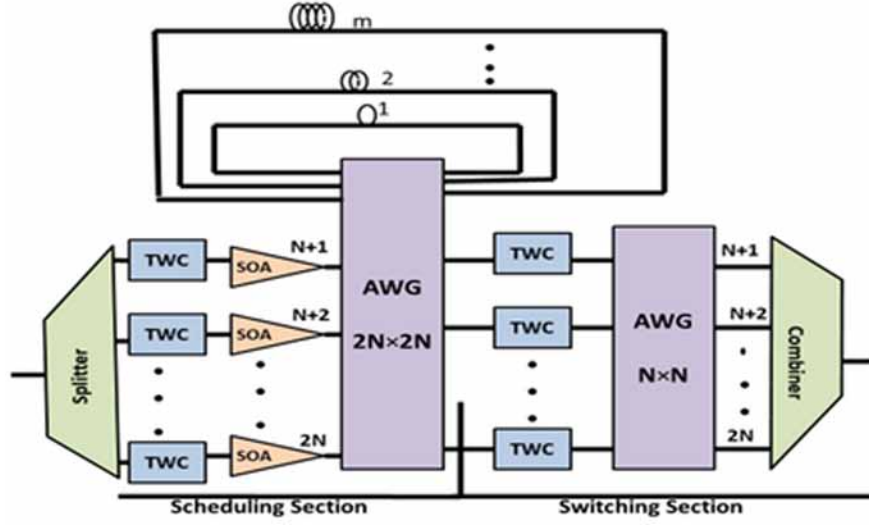
PERFORMANCE OF SWITCH IN TO THE NETWORK

In the previous sections the performance of switch is measured as an isolated switch. In this section the overall performance of switch is considered when it is placed in to the network where cascading of multiple switches is required. Figure 11 shows the structure of switch when it is placed in to the network where multiplexed data traverses from one node to other. The multiplexed data will arrive at the switch input then first the input data is first demultiplexed however, as the data can arrive at any wavelength range, splitter should be used at the input, and combiner at the switch output. The structure of switch is presented as in figure 12.

THE LOSS ANALYSIS

The loss analysis of switch when packet passes through the longest FDL of buffer. The splitter and combiner both are lossy devices and their insertion loss is high in comparison to other devices so use of splitter and combiner increases the overall loss of the switch. The overall loss of switch is calculated as follows by equation (15).

Figure 11. Switch design ready to use in the network (D3).



$$L(S) = L_{Splitter} + L_{TWC} + L_{AWG} + L_{FDL} + L_{TWC} + L_{AWG} + L_{SOA} + L_{Combiner} \quad (15)$$

Now putting different loss values the overall loss of the switch is given by eq. (16)

$$L(S) = 12 + 2 + 3 + 1 + 2 + 3 + 1 + 12 = 36dB \quad (16)$$

In above calculation the size of scheduling AWG is considered as 8×8 . Hence when data traverse through the switch, it faces huge insertion loss of 32dB. So, in the network, for compensating this huge loss optical amplifier like EDFA is required. Beyond this, the loss is also minimized by replacing splitter and combiner by Demux and Mux. The combination of Demux and Mux do the same job as splitter and combiner with very less insertion loss. After this replacement the total loss of the switch is given by eq. (17)

$$L(S) = L_{Demux} + L_{TWC} + L_{AWG} + L_{FDL} + L_{TWC} + L_{AWG} + L_{SOA} + L_{Mux} \quad (17)$$

By putting the various loss values the total loss is calculated in eq. (18):

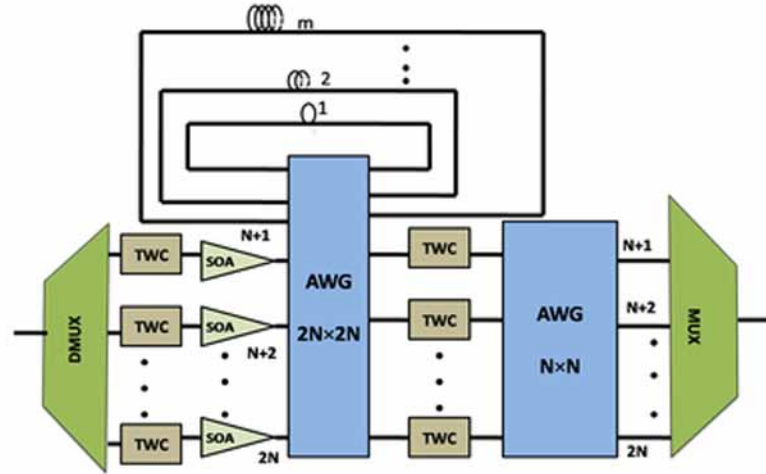
$$L(S) = 2.5 + 2 + 3 + 1 + 2 + 3 + 1 + 2.5 = 17dB \quad (18)$$

After replacing Splitter and combiner by Mux and Demux the new design of switch, ready to use in the network is as shown in the figure 12.

Power Analysis of Switch D4

Using, the analysis as done in previous sections of this paper, power at the output of the switch D4 for bit b is

Figure 12. Switch design ready to use in the network (D4)



$$P_{out} = bP_{in} + n_{sp}(G-1)h\nu B_0 L_{AWG}^{2N \times 2N} L_{TWC} L_{AWG}^{N \times N} L_{MUX} \quad (19)$$

In the previous sections of this paper analysis of an isolated switch design is presented. When switch is placed inside the network switch D4 emerges.

Let us consider a network as shown in the figure 13. The network presented in to the graph has total of 6 nodes in to the networks distances among different nodes are shown through the edges between these nodes. In the directed graph the node 1 is the source node while node 4 acts as the destination node. Now considering the shortest path among source and destination node in the graph the shortest path from source to destination is 1-6-3-4 and shortest distance from source to destination is 1.64 unit.

In general, distances among adjacent nodes in optical core networks are in some hundreds to some thousands of kilometres. Considering 1unit=1000km. So, the total travelled distance in the graph is 1640 km. Let us suppose the loss of the fiber is 0.2dB/km. So, the total calculated loss is 328dB. Suppose EDFA of gain 30dB each are placed in the links of the network. In this case for compensating losses of 328 dB the required numbers of amplifiers are 11 (10 of 30 dB and one of 28 dB). Similarly, we can count the number of amplifiers required for any of the network.

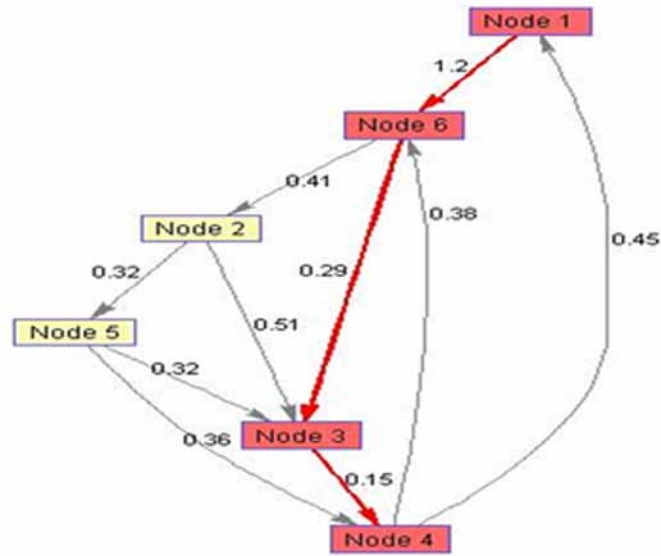
In network, the new equation for power calculation would be given by eq. (20)

$$P_{out} = bP_{in} (L_{in} G_{in})^K + N_{sd} n_{sp} (G-1)h\nu B_0 L_{AWG}^{2N \times 2N} L_{TWC} L_{AWG}^{N \times N} L_{Mux} + K n_{sp} (G_{in} - 1)h\nu B_0 \quad (20)$$

Where, L_{in} and G_{in} are the loss and gain of the inline amplifier, and 'K' is the number of inline amplifiers and N_{sd} is the numbers of switch through which data moves between source to destination. Again, by using equation (20) and choosing the values of different parameters from table 2, BER can be re-evaluated at different power levels.

In the previous sections, from the analysis presented for switch D2 the BER analysis at various power levels is presented in Table 3 to 5 for different switch configurations. BER vs. power is shown as isolated switch in Table 3. It is obvious that as the power increases BER improves. For acceptable $BER \leq 10^{-9}$, the minimum amount of required power is 140 nW which is very less in comparison to the earlier switch

Figure 13. Considered network Digraph.



designs where power requirements are in micro and sub-micro levels. When switch is made compatible with network, the power requirement increases to 160 nW (Table 4).

Table 3. Bit error rate for switch D2 at various power levels.

Power in nW	BER
100	4.38×10^{-7}
110	6.11×10^{-8}
120	8.07×10^{-9}
130	1.01×10^{-9}
140	1.21×10^{-10}
190	1.68×10^{-15}
200	1.62×10^{-16}

Table 4. Bit error rate for switch D3 at various power levels.

Power in nW	BER
120	4.14×10^{-7}
130	7.51×10^{-8}
140	1.30×10^{-8}
150	2.15×10^{-9}
160	3.41×10^{-10}
200	1.51×10^{-13}

Table 5. Bit error rate for switch D4 at various power levels when it is placed in to network.

Power in μW	BER
1.3	3.56×10^{-8}
1.4	9.11×10^{-9}
1.5	2.31×10^{-9}
1.6	5.86×10^{-10}
1.7	1.47×10^{-11}
1.8	3.69×10^{-11}
1.9	9.22×10^{-12}
2.0	2.29×10^{-12}

When switch placed in the network, the power requirement is 1.6 micro-watts (Table 5). This analysis is very meaningful, as in optical technology, 3R regeneration is not possible, and so from the transmitted node packet should be generated with the required amount of power so it can reach its destination.

Table 6. Comparative analysis of switch.

Switch Design	Power
Isolated Switch (D2)	130 nano-watts
Switch ready to place in Network (D3)	160 nano-watts
Switch in the network (D4)	1.6 μW

Table 6, compares the switch design and acceptable power levels. It is obvious from the table that there is huge power difference in different configurations of the switch. It must be remembered that the required amount of power will vary network to network as both distance between source to destination node pair will change and number of hops traverse will also change.

FUTURE RESEARCH DIRECTIONS

The one of the major challenges in the deployment of OPS is that the transferring of packet only in the optical domain without O-E-O conversion is not possible. Electronic packet switching employs input and output buffering to perform store and forward processing needed at each switching node. This is not possible just because the unavailability of optical RAM. Two approaches are commonly used to overcome the lack of optical RAM, we can use fiber delay lines-based buffers that stores packets in to optical domain. The switch discussed in to the article is FDL based switch and it uses fiber delay lines for storing the packets whenever situation of contention among packets arises. The major drawback of using FDL lines are that, this approach significantly increases the complexity of the switch architecture and its control. The extra latency is also introduced in to the system. Moreover, this method affects the signal power and by this method the packets are delayed only for fixed time, so the method does not solve the problem completely. The second approach which is used in the OPS switches are to introduces buffering at the edges of switching network. In this method buffering can be implemented by using the

electronic buffer at the edge while on the other hand switching process is in optical domain. But this method is applicable only in the systems that feature short host switch distances. In fact, in the situation of contention, the packets which are to be stored at the edge of switch needs to be retransmitted. In the present day's optical environment, the researchers are focused in the design of hybrid buffer which contains all the quality features of electronic as well as optical buffering. Some of the switches using hybrid buffer are presented in by P. Nair and Ramesh P.G.V. (2013). This paper presents a scheme for hybrid burst assembly and a hybrid burst loss recovery scheme which selectively employs proactive loss recovery mechanism depending on the classification of traffic into short term and long term, respectively. Traffic prediction is made using the hidden Markov model (HMM). This dynamic hybrid approach based on HMM prediction provides overall a lower blocking probability and delay and more throughput when compared with forward segment redundancy mechanism. Ahmed Nabih et al. (2014) a hybrid optical switch (HOS) based on wavelength division multiplexing and optical code division multiplexing (WDM/OCDM) scheme is detailed. In this optical cross connect (OXC) can control and process requests for both circuit establishment and burst scheduling. The results are presented in terms of the blocking probability, probability of error, and probability of outage. Except these switches the numbers of other switches are also available that uses hybrid buffering techniques. So, from this study we can say that the scope of research in the design of switch are not limited and there are number of possibilities, from this article we can analyse the effect of each optical component on the performance of switch which gives a brief layout of switch design.

CONCLUSION

In this paper the performance analysis of an AWG based switch is presented, this switch is realized by using TWC, AWG and piece of fiber in the core of switch. In this architecture the routing pattern of AWG is considered and its symmetric nature, simplify switch operations significantly. In this paper design modifications are suggested and a comparison between original design and modified designs are performed through physical layer analysis of each switch. The obtained results presented in this paper clearly favour the suggested modifications.

- The total physical loss of modified switch is fully compensated by corresponding SOA presented at each input of switch both buffer and straight through losses are fully compensated by SOA. This makes modification more useful.
- The obtained results presented in this paper clearly reveals that the nano-watt power level is required in the correct operations of switch D2, which is very less in comparison to other switches available in to market.
- In the analysis presented in the paper obtained results clearly reveals that analysis is meaningful in the absence of 3R and when cross-layer optimization is done. It has been found that as in isolated switch power requirement is 140 nano watts which increase to 1.6 microwatts in the considered network.

REFERENCES

- Habib, M. F., Tornatore, M., De Leenheer, M., Dikbiyik, F., & Mukherjee, B. (2012). Design of disaster-resilient optical datacenter networks. *Journal of Lightwave Technology*, 30(16), 2563–2573. doi:10.1109/JLT.2012.2201696
- Koley, B., Vusirikala, V., Lam, C., & Gill, V. (2010, July). 100GbE and beyond for warehouse scale computing. In *OECC 2010 Technical Digest* (pp. 106–107). IEEE.
- Katz, R. H. (2009). Tech titans building boom. *IEEE Spectrum*, 46(2), 40–54. doi:10.1109/MSPEC.2009.4768855
- Sato, K. I., & Hasegawa, H. (2009). Optical networking technologies that will create future bandwidth-abundant networks. *Journal of Optical Communications and Networking*, 1(2), A81–A93. doi:10.1364/JOCN.1.000A81
- Agrawal, G. P. (2012). *Fiber-optic communication systems* (Vol. 222). John Wiley & Sons.
- Koch, T. L. (2012). *Optical Fiber Telecommunications IIIB*. Elsevier.
- Tucker, R. S. (2008). Optical packet switching: A reality check. *Optical Switching and Networking*, 5(1), 2–9. doi:10.1016/j.osn.2007.08.001
- Yoo, S. B. (2006). Optical packet and burst switching technologies for the future photonic internet. *Journal of Lightwave Technology*, 24(12), 4468–4492. doi:10.1109/JLT.2006.886060
- Xu, L., Perros, H. G., & Rouskas, G. (2001). Techniques for optical packet switching and optical burst switching. *IEEE Communications Magazine*, 39(1), 136–142. doi:10.1109/35.894388
- Mujherjee, B. (n.d.). *Optical WDM Networks*. Academic Press.
- Stern, T. E., & Bala, K. (1999). *Multiwavelength optical networks: a layered approach*. Addison-Wesley Longman Publishing Co., Inc.
- Senior, J. M. (2004). *Optical fiber communications: principles and practice*. Academic Press.
- Qiao, C. (2000). Optical Burst Switching (OBS)-A new paradigm for an optical internet. *J. High Speed Networks*, 8(1), 36–44.
- Masetti, F., Zriny, D., Verchere, D., Blanton, J., Kim, T., Talley, J., & Poignant, P. (2002, March). Design and implementation of a multi-terabit optical burst/packet router prototype. In *Optical Fiber Communication Conference and Exhibit* (pp. FD1–FD1). IEEE. doi:10.1109/OFC.2002.1036782
- Turner, J. S. (1999). Terabit burst switching. *Journal of High-Speed Networks*, 8(1), 3–16.
- Chen, Y., Qiao, C., & Yu, X. (2004). Optical burst switching: A new area in optical networking research. *IEEE Network*, 18(3), 16–23. doi:10.1109/MNET.2004.1301018
- Kato, K., Okada, A., Sakai, Y., Noguchi, K., Sakamoto, T., Suzuki, S., Takahara, A., Kamei, S., Kaneko, A., & Matsuoka, M. (2000). 32× 32 full-mesh (1024 path) wavelength-routing WDM network based on uniform-loss cyclic-frequency arrayed-waveguide grating. *Electronics Letters*, 36(15), 1294–1296. doi:10.1049/el:20000916

- Ramaswami, R., Sivarajan, K., & Sasaki, G. (2009). *Optical networks: a practical perspective*. Morgan Kaufmann.
- Rastegarfar, H., Leon-Garcia, A., LaRochelle, S., & Rusch, L. A. (2012). Cross-layer performance analysis of recirculation buffers for optical data centers. *Journal of Lightwave Technology*, 31(3), 432–445. doi:10.1109/JLT.2012.2226865
- Yin, Y., Proietti, R., Ye, X., Nitta, C. J., Akella, V., & Yoo, S. J. B. (2012). LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers. *IEEE Journal of Selected Topics in Quantum Electronics*, 19(2), 3600409–3600409. doi:10.1109/JSTQE.2012.2209174
- Bhattacharya, P., Singh, A., Kumar, A., Tiwari, A. K., & Srivastava, R. (2017, November). Comparative study for proposed algorithm for all-optical network with negative acknowledgement (AO-NACK). In *Proceedings of the 7th International Conference on Computer and Communication Technology* (pp. 47–51). ACM. 10.1145/3154979.3154981
- Singh, A., Tiwari, A. K., & Srivastava, R. (2018). Design and analysis of hybrid optical and electronic buffer based optical packet switch. *Sadhana*, 43(2), 19. doi:10.1007/12046-018-0786-1
- Shukla, V., & Srivastava, R. (2015). WDM fiber delay lines and AWG based optical packet switch architecture. *Proceedings of National Conference on Innovative Trends in Computer Science Engineering (ITCSE-2015)*.
- Srivastava, R., & Singh, Y. N. (2010). Feedback fiber delay lines and AWG based optical packet switch architecture. *Optical Switching and Networking*, 7(2), 75–84. doi:10.1016/j.osn.2010.01.002
- Srivastava, R., Singh, R. K., & Singh, Y. N. (2008). WDM-based optical packet switch architectures. *Journal of Optical Networking*, 7(1), 94–105. doi:10.1364/JON.7.000094
- Shukla, V., Sonkar, R. K., & Srivastava, R. (2019). Performance comparison between recirculating loop buffer-based optical packet switch architectures. *Indian Journal of Physics*, 1–12.
- Ramesh, P. G. V., & Nair, P. (2013). A multi-layer approach for load balancing in optical burst switching networks. *Optik (Stuttgart)*, 124(17), 2602–2607. doi:10.1016/j.ijleo.2012.07.037
- Rashed, A. N. Z. (2014). High efficiency wireless optical links in high transmission speed wireless optical communication networks. *International Journal of Communication Systems*, 27(11), 3416–3429.

KEY TERMS AND DEFINITIONS

Amplifier: The amplifiers are the devices which are used to amplify the signals when the signal strength goes down.

AWG: The AWG router is used to map an input waveguide to an array of output waveguides in such a way that the different wavelength signals present in the input waveguide are mapped on the different output waveguides.

BER: As the name implies the bit error rate is defined as the rate at which the error occurs in the transmission medium.

Fiber Delay Lines: The fiber delay lines (FDL) are used in the switch for storing the packets in to the optical domain.

OPS: In optical networks, it is assumed that the data traverse from one node to another node in a form of fixed size packets are referred as optical packet switching (OPS).

RAM: Due to unavailability of optical storage device the RAM is used as the electronic buffer in to the design of switch.

WDM: The wavelength division multiplexing is a technique which modulate the optical carrier signals of varying wavelengths of laser light, onto a single optical fiber.

Formal Verification of ZigBee–Based Routing Protocol for Smart Grids

Adnan Rashid

 <https://orcid.org/0000-0002-9729-584X>

National University of Sciences and Technology, Pakistan

Osman Hasan

 <https://orcid.org/0000-0003-2562-2669>

National University of Sciences and Technology, Pakistan

INTRODUCTION

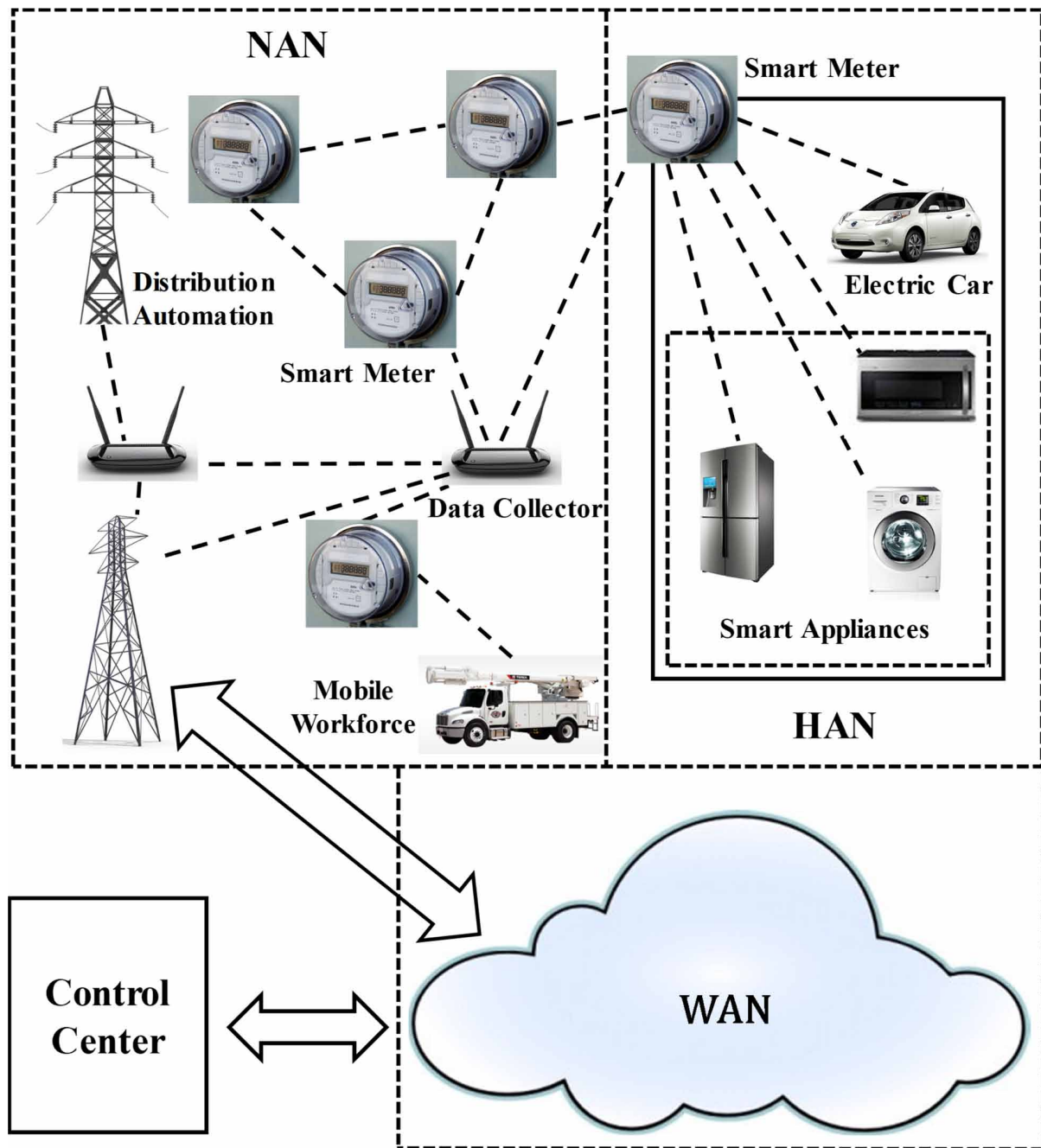
With the growing demand of electricity, the traditional power transmission grids are unable to manage instant upsurges in the utilization of electricity and thus often lead to blackouts and power outages (Gao, Xiao, Liu, Liang & Cheng, 2012). Smart grid technology (Farhangi, 2010), which is a digital upgradation of the traditional grids, can overcome the above-mentioned issues and thus provides reliable, secure, safe and efficient power distribution and management. A smart grid system is mainly composed of control centers, mobile workforce, substations, utilities and the customer premises (Saputro, Akkaya & Uludag, 2012). A reliable communication link between all these building blocks is the core element that governs the efficiency of the overall system. Some of the extensively used communication networks in this technology are wide area network (WAN), neighborhood area network (NAN) and home area network (HAN) (Saputro et al., 2012). For example, HAN is located in the customer premises and thus provides a connection between the smart meter and the home appliances. These networks and their interrelationships are depicted in Figure 1.

The routing protocols play a vital role in establishing a secure and reliable communication link between the components of the above-mentioned smart grid networks. In this regard, the ZigBee (Kinney et al., 2003) routing protocol is commonly used in wireless HAN and provides a reliable, secure and uninterrupted connection between various components of HAN, i.e., home appliances and the smart meter (Saputro et al., 2012). Conventionally, the analysis of these network protocols is performed using live testing and computer-based simulations (Lundgren, 2002) using network simulators (McCanne, Floyd, Fall & Varadhan, 1997). These techniques can analyze a limited number of possible scenarios and thus there is always a chance of missing a corner case, which compromises the accuracy of the analysis. Moreover, round-off errors due to the involvement of the computer arithmetic also introduce some approximations and thus we cannot certify a complete absence of bugs. Therefore, considering the safety-critical applications of power and electricity in our daily tasks, these inaccurate results may lead to many unwanted consequences.

Formal methods (Hasan & Tahar, 2015) have the ability to overcome these inherent limitations of the above-mentioned conventional techniques. The formal methods based system analysis mainly involves developing mathematical model of the given system based on some logic and verification of its intended properties using deductive reasoning. The involvement of the logical modeling and deductive reasoning enhances the chances of catching subtle errors that are often ignored by the traditional methods. For

DOI: 10.4018/978-1-7998-3479-3.ch069

Figure 1. Smart grid Communication Network



example, for a 32-input adder, we require $2^{64} = 1.8446744 \times 10^{19}$ test vectors in the case of computer-based simulations and it is checked for a limited number of these test vectors to reduce the number of computational overheads. Thus, there is always a chance of missing a bug due to not using the test vector to catch it. However, in the case of formal methods, we need to develop an implementation and specification of the circuit using some appropriate logic and verify that the implementation meets the

specifications of the underlying system. The two most commonly used formal methods include model checking (Baier & Katoen, 2008) and theorem proving (Harrison, 2009).

Model checking is an automated analysis technique, which involves the state-space based modeling of the underlying system and rigorously verifying its properties specified in an appropriate logic. The discrete-time systems are commonly verified using this technique due to its state-based nature. Moreover, it has been frequently used for analyzing concurrent systems (Clarke, 1986; Gradara, 2007). However, model checking is not very suitable for systems exhibiting the continuous behavior due to its inherent state-space explosion problem (Baier & Katoen, 2008), i.e., the issues related to the limited amount of computer memory and resources for larger systems.

On the other hand, theorem proving is a formal analysis technique, which involves developing a mathematical model of the system to be analyzed in an appropriate logic and verification of the properties of interest using computer based tools known as theorem provers/proof assistants. Based on the decidability or undecidability of the underlying logic, theorem proving can be termed as automated or interactive. Propositional logic is decidable and thus the sentences/formulas expressed in this logic can be verified automatically. However, one of its main limitations is less expressiveness and thus it cannot be utilized for analyzing all sort of systems. Higher-order logic is the most expressive form of logic and it allows modeling and verifying the complex systems exhibiting the continuous behavior. However, due to the undecidable nature, the verification of theorems about sentences expressed in higher-order logic requires a lot of user guidance in an interactive manner.

Due to the involvement of a lot of human interaction, theorem proving is not preferred for the systems exhibiting the discrete-time behavior. However, the automatic nature of model checking along with the suitability for formally analyzing the concurrent systems makes it an ideal candidate for the verification of network and routing protocols. This is one of the major factors for choosing model checking for the formal verification of the ZigBee-based routing protocol for smart grids.

In this chapter, we use the UPPAAL model checker (Behrmann, David & Larsen, 2004) to formally verify the ZigBee protocol for smart grids. The primary motivation for choosing UPPAAL is that it can handle time in its true form (as clock variable) and the fact that it has been used for modeling and the automatic verification of many systems, in particular the network and routing protocols. To illustrate the effectiveness of our idea, we use our UPPAAL model to verify the collision avoidance, bounded liveness and deadlock freeness properties of the smart grids' ZigBee-based protocol.

RELATED WORK

Model checking has been widely used to verify a variety of communication protocols for smart grids. Fehnker et al. (2007) formally verified a light-weight medium access control (LMAC) protocol, which operates in multi-hop energy-constrained wireless sensor networks (WSNs). Through this exercise, the authors mainly detected and resolved the collision between the data packets using UPPAAL model checker. Similarly, a formal modeling and verification of the ad hoc on-demand distance vector (AODV) protocol operating in wireless mesh network (WMN) is presented in Fehnker et al. (2012). The authors used the algebra for wireless networks (AWN) process algebra to model the AODV routing protocol. Next, this AWN model is converted into its equivalent UPPAAL model, which is further used for verifying its properties based on its different dynamic topologies. Hofner et al. used statistical model checker SMC-UPPAAL for the formal quantitative analysis of AODV and dynamic MANET on-demand (DYMO) protocols for WMNs (Hofner & McIver, 2013).

Malik et al. (2013) proposed a framework for the formal analysis of inter control center communications protocol (ICCP) used in exchanging data and control in different control centers. The authors modeled the protocol using an iterative process and verified its safety and critical security properties using the UPPAAL model checker. Similarly, Tschirner et al. (2008) used UPPAAL to verify the quality-of-service (QoS) of biomedical sensor networks (BSNs). The authors mainly verified the network connectivity, packet delivery and deadlock freeness properties using their UPPAAL model. Somappa et al. (2015) provided the verification of the dual-mode adaptive MAC (DMAMAC) protocol, which is used in process control applications. The authors used UPPAAL to formally verify the state-switching and deadlock freeness properties of the DMAMAC protocol. Similarly, Rashid et al. presented the formal analysis of the ZigBee routing protocol using UPPAAL (Rashid, Hasan & Saghar, 2015) for smart grid applications. The context of this book chapter is mainly focused on this work as, to the best of our knowledge, this is the only work that is related to the verification of communication protocols for smart grids.

Kwiatkowska et al. (2002) used probabilistic model checking to formally analyze the IEEE 802.11 wireless local area network (WLAN) protocol using a two step process. The first step involves the construction of a model based on probabilistic timed automaton. In the next step, the property-preserving semantics are used to convert this automaton model into a finite-state Markov decision process, which is verified using the PRISM model checker. Similarly, Ballarini & Miller (2006) used PRISM to formally analyze the sensor-MAC (S-MAC) protocol for WSN with respect to energy consumption and performance.

Fehnker et al. (2009) proposed a framework to analyze the wireless network protocols, which is based on both the simulation and model checking techniques. The authors first modeled the wireless protocol using CaVi, which is a graphical user interface used for the modeling of networks. In the next step, this CaVi model is translated to the Castalia simulator and PRISM models. Finally, the experiments are performed on the simulator to judge the performance of the protocol. Also, the performance properties of the protocol are verified using the PRISM model checker. Similarly, Rossi et al. (2013) presented formal verification and performance analysis of the cognitive wireless networks based on their discrete time Markov chain (DTMC) models using PRISM.

Bhargavan et al. (2002) formally verified the AODV and routing information protocol (RIP) using the SPIN model checker along with the HOL theorem prover. Similarly, Renesse & Aghvami (2004) used SPIN to verify wireless ad-hoc routing protocol (WARP). Islam et al. (2006) formally modeled the dynamic host configuration protocol (DHCP) using process meta language (PROMELA) and formally verified its model using SPIN. A detailed account of the automated formal analysis of routing protocols for WSNs can be found in Chen et al. (2013). Moreover, a survey of the formal analysis of communication network protocols can be found in Qadir & Hasan, (2015).

The area of formal verification of communication network and routing protocols for smart grid systems has not been fully explored yet. There is only a single work reported in the literature, in which the authors presented the formal analysis of the ZigBee routing protocol for smart grids using UPPAAL (Rashid et al., 2015). Therefore, we have chosen to focus on this work in this chapter. Moreover, we have further elaborated the work, which ensures its better understanding to the non-formal methods community and thus addresses a wider audience.

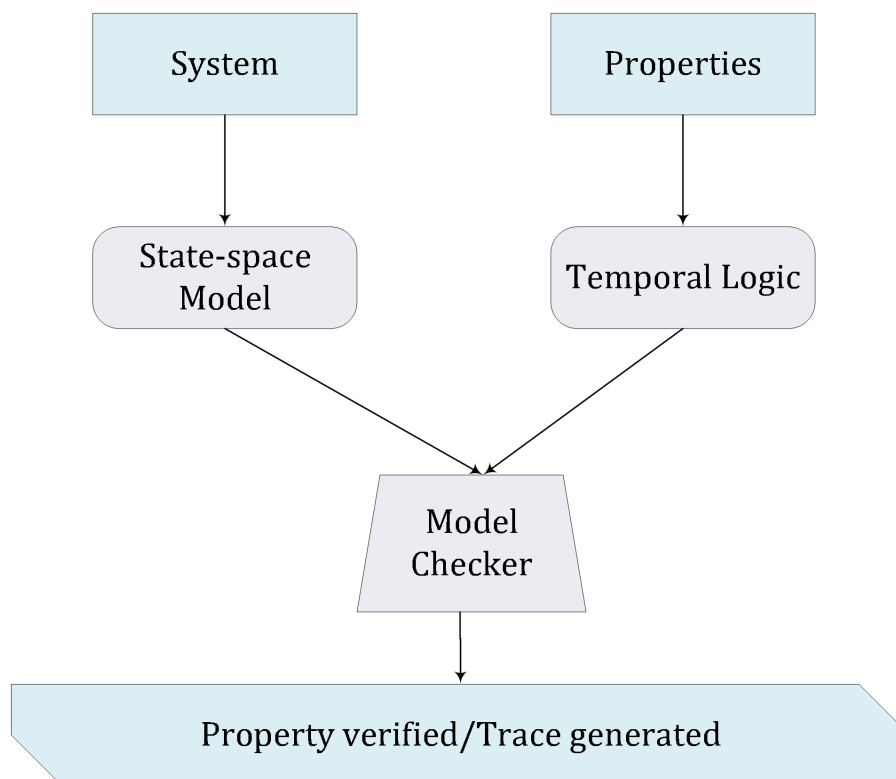
BACKGROUND

In this section, we briefly describe model checking and the UPPAAL model checker to facilitate the understanding of the formal modeling and verification of the ZigBee-based routing protocol presented later in the chapter.

Model Checking

Model checking (Clarke et al., 1999) is one of the major formal techniques, which is commonly used for analyzing concurrent systems, such as network and routing protocols. The model checking-based analysis involves the construction of a state-space model of the given system and specification of the intended properties of the system in temporal logic (Pnueli, 1977), as depicted in Figure 2. Next, both the model and properties are given to the model checker, which explores the state-space exhaustively and automatically verifies the given system based on these properties. In the case of failure of a property, the model checker generates an error-trace, which helps a user to identify and rectify the error found in the system's model. For the larger systems, this technique is subject to the problem of state-space explosion (Baier & Katoen, 2008) caused by the limited availability of computer memory and other computational resources. However, the abstraction of the model or usage of the efficient bounded and symbolic model checking techniques can avoid this problem.

Figure 2. Model Checking



UPPAAL Model Checker

UPPAAL (Behrmann et al., 2004) model checker is extensively used to formally analyze real-time systems. The behavior of the given system is represented as a network of timed automata. The synchronization between each of the individual automaton is done using the phenomenon of channels, which are of two types: namely binary and broadcast synchronization channels. The binary channel is used for synchronizing one automaton with the other, whereas, in the case of broadcast channel, a single automaton is synchronized with all the channels with enabled transitions using *committed locations* in UPPAAL. These modeling and verification features enable UPPAAL as a commonly used tool to formally analyze concurrent systems, such as communication network and routing protocols (Yi, Pettersson & Daniels, 1994).

We can formally verify the UPPAAL model of a system using the properties specified in computational tree logic (CTL) (Orgun & Ma, 1994), which are the formulas in temporal logic and support specifications based on temporal, path and logical operators. The temporal operators involve next (X), always ([]) and eventually (< >). Similarly, the path operators are thereexists (E) and forall (A). Whereas, negation (!), disjunction (||), conjunction (&&) and implication (@) are the logical operators. Table 1 presents the CTL property specification and their graphical representation.

ZIGBEE ROUTING PROTOCOL

ZigBee (Saputro et al., 2012) is a technology, which resides at the top of networking standard IEEE 802.15.4 and is widely used in HANs for smart grids. It exhibits many interesting characteristics, like low power, long battery life and low data rates. These distinguishing features make it a suitable candidate to operate in short range frequencies. In the HAN, ZigBee is mainly responsible for an uninterrupted communication between various smart appliances, i.e., lightings, air conditioner, refrigerator, oven and fans etc. It is build up of four layers that are stacked over each other in the network and include MAC, physical, network and application layers. IEEE 802.15.4 standard is used to define the physical and MAC layers, whereas the application and network layers are based on ZigBee specification (Gomez & Paradells, 2010). ZigBee-based routing protocol uses the master-slave strategy for routing purposes and its architecture is depicted in Figure 3. Here, the ZigBee coordinator is a master and the slaves include various home appliances, i.e., oven, television and refrigerator etc (Hafeez, Kandil, Al-Omar, Landolsi & Al-Ali, 2014). The master (ZigBee coordinator) is mainly responsible for establishing a reliable and uninterrupted communication in the network and thus it directly communicates with all slaves (home appliances) and also manages their intercommunication by coordinating with the home appliances.

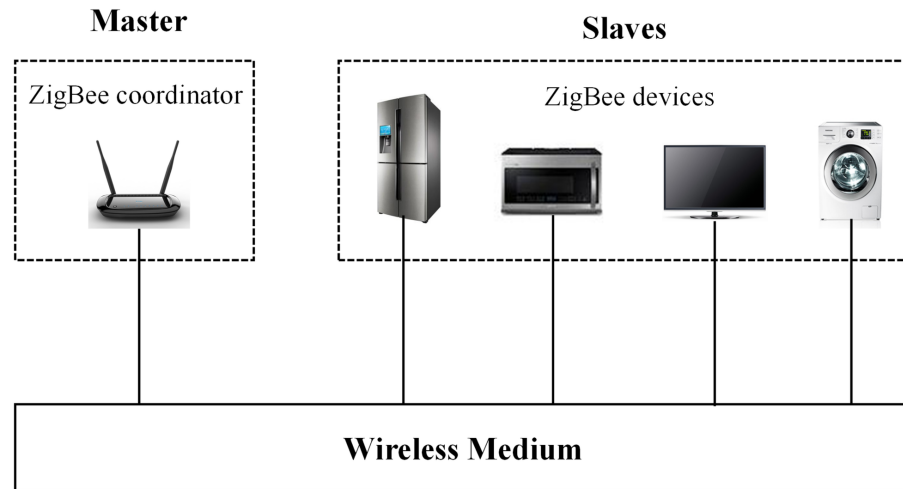
The ZigBee-based routing protocol provides the following advantages in HANs:

1. It consumes low power and helps the batteries to last longer for the ZigBee devices present in the network.
2. It helps in avoiding conflicts and collision between different nodes by providing a reliable and highly secured network.
3. It has a very low latency, i.e., about 15ms to 30ms.
4. It has abilities of self-organization in a network (Huq & Islam, 2010).

Table 1. Property Specification in Computational Tree Logic (CTL)

CTL Property	Description	Graphical Representation
$E < > p$	There exists a path where the property p eventually holds	
$A [] p$	p always holds for all the paths	
$E [] p$	There exists a path where p always holds	
$A < > p$	p will eventually hold for all paths	
$P \textcircled{R} q$	Whenever the property p holds q will eventually hold	

Figure 3. Architecture of Zigbee Protocol



FORMAL MODEL FOR ZIGBEE-BASED ROUTING PROTOCOL

The ZigBee-based routing protocol uses the master-slave strategy and thus the main components of its real-time UPPAAL model are the *master* and *slaves* automata modeling the ZigBee coordinator and smart appliances, respectively. The master broadcasts the data to slaves via a wireless channel. Therefore, we require a *medium* automaton that models the channel in our UPPAAL model. The data sending and receiving abilities of a slave are modeled using the *user* automaton. Finally, to model the time bound in the data transmission, we use a *test* automaton, which is later used to verify the bounded liveness property of the underlying system. We consider the following assumptions for our UPPAAL model:

1. Lossless medium for the communication between various components, i.e., every data sent on the channel will surely be received.
2. Two slaves are considered for our UPPAAL model.

The timed automaton corresponding to the master process is depicted in Figure 4. The master automaton initiates the process by sending an enquiry to the first slave, which is represented by the initial transition, i.e., *mas_1* to *mas_2*. The data (enquiry/message) that needs to be transmitted, is modeled by the variable *data* and an enquiry is represented by *data:=0*. Similarly, the update *sla_num:=1* represents the first slave. Now, the channel synchronization *med_free?* is used in the next transition to ensure that the master waits until the data is broadcasted to all the slaves. In the third transition, the broadcasted data from a slave is received by the master. We also specified a condition $t_mas \leq 3$ working as a clock invariant in the master state *mas_3* to ensure that the data is received within 3 time units, which actually models the assumption of the lossless medium. Finally, an increment in the variable *sla_num* in transition from *mas_4* to *mas_2* shows that an enquiry is sent to next slave from the master automaton.

The medium automaton is depicted in Figure 5. The data sent by the master is received by the medium process by synchronizing the input channel *sen_med?* with the master. Next, the medium process broadcasts the received data to all the slaves using the phenomenon of *committed locations*, which enables the broadcast feature for the UPPAAL model. In the case, if a collision is occurred on the medium, the given system will move to *coll_state*.

Figure 4. The Master Automaton

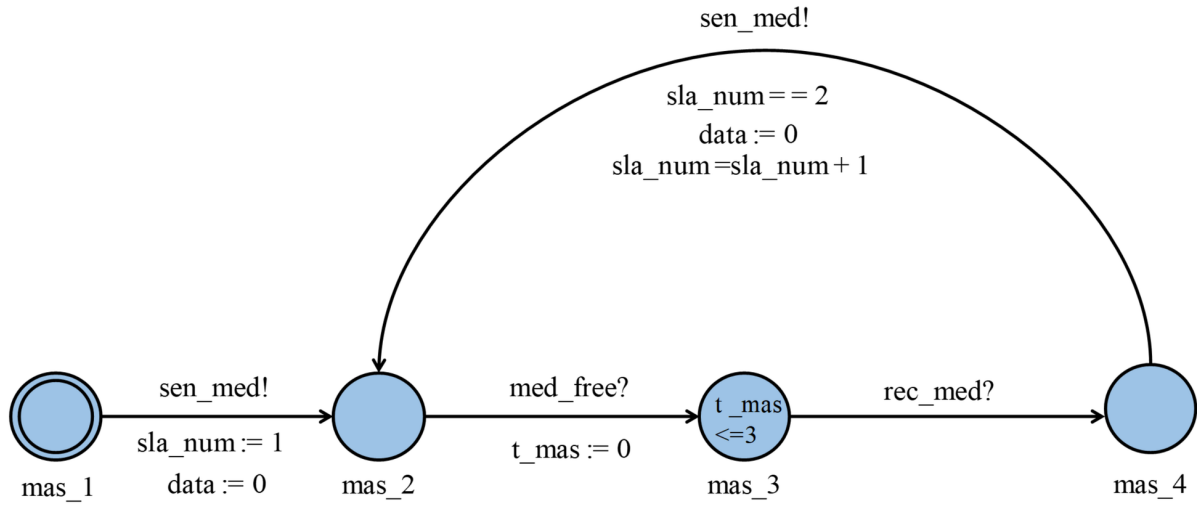
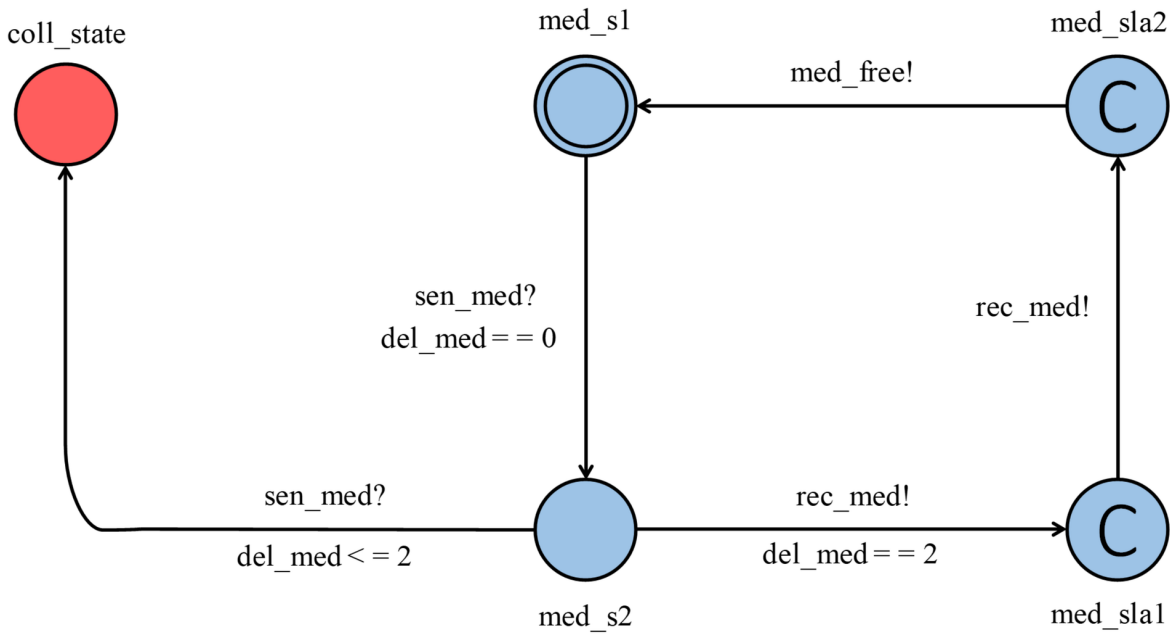
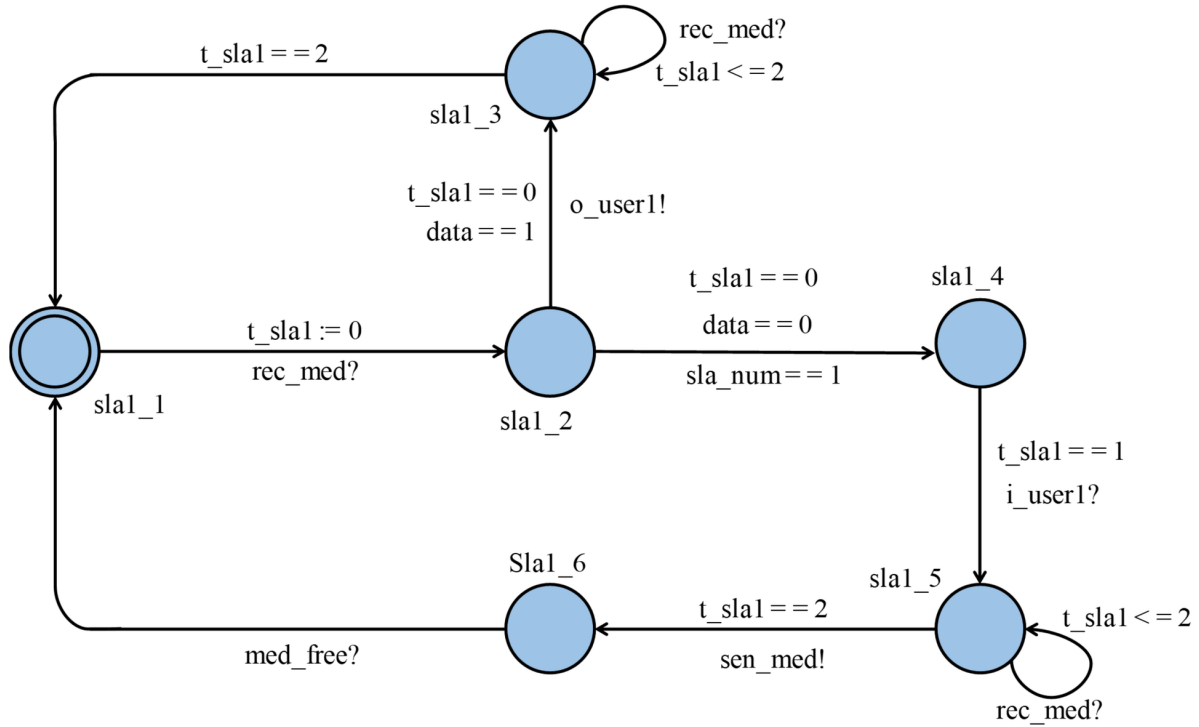


Figure 5. The Medium Automaton



We modeled both the slaves as individual automata, which are almost same, except a couple of different parameters. Figure 6 depicts the automaton for the first slave process. The slave automaton initiates the process from sla1_1 and receives data as a result of synchronization with the medium using input action `rec_med?`, which is represented by the transition from sla1_1 to sla1_2. Now, the slave automaton at location sla1_2 has two possible transitions. The slave either sends an enquiry to the user by transitioning from sla1_2 to sla1_3 or sends a message by transitioning from sla1_2 to sla1_4. For both the cases, the slave will not answer to any user for a couple of time units, which is illustrated by the self-transitions on sla1_3 and sla1_5, respectively.

Figure 6. The First Slave Automaton



A user automaton is modeled for each of the slave processes, which captures the capabilities of data transmission. Figure 7 depicts the first slave's user automaton. It uses the output actions `sen_u1` and `rec_u1` for the data transmission.

We used a test automaton, as depicted in Figure 8, to verify the bounded liveness of the ZigBee protocol. It utilizes a clock variable `t_tes` to express the time bound in the transmission of data. If the transmitted data is not received within a certain time limit, the test automaton will enter into location `sen_failed` for the case when data is not sent within time limit. In the case, if the data is sent successfully and not received then it will be forced into `rec_failed`.

FORMAL VERIFICATION OF ZIGBEE-BASED ROUTING PROTOCOL

As the ZigBee-based routing protocol works on the master-slave architecture, it ensures that there is no collision during the data transmission from master to slaves or one slave to the other slave. Moreover, the sent data should be received within a time limit. Both of these properties authenticate our design and formal model based on UPPAAL model checker. We used the following CTL formula to verify the collision avoidance property:

$$\forall [] (\text{not medium.coll_state})$$

Figure 7. First Slave's User Automaton

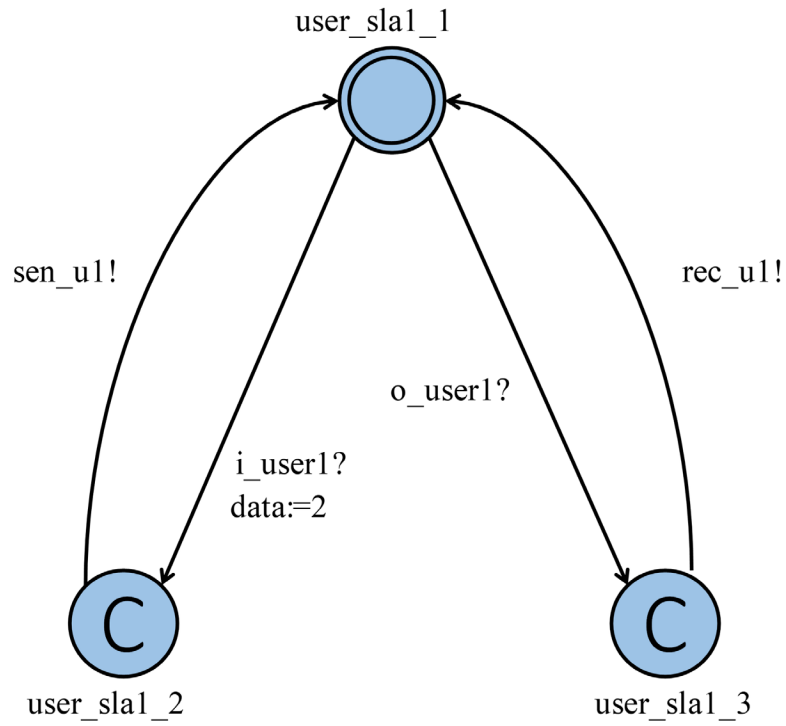
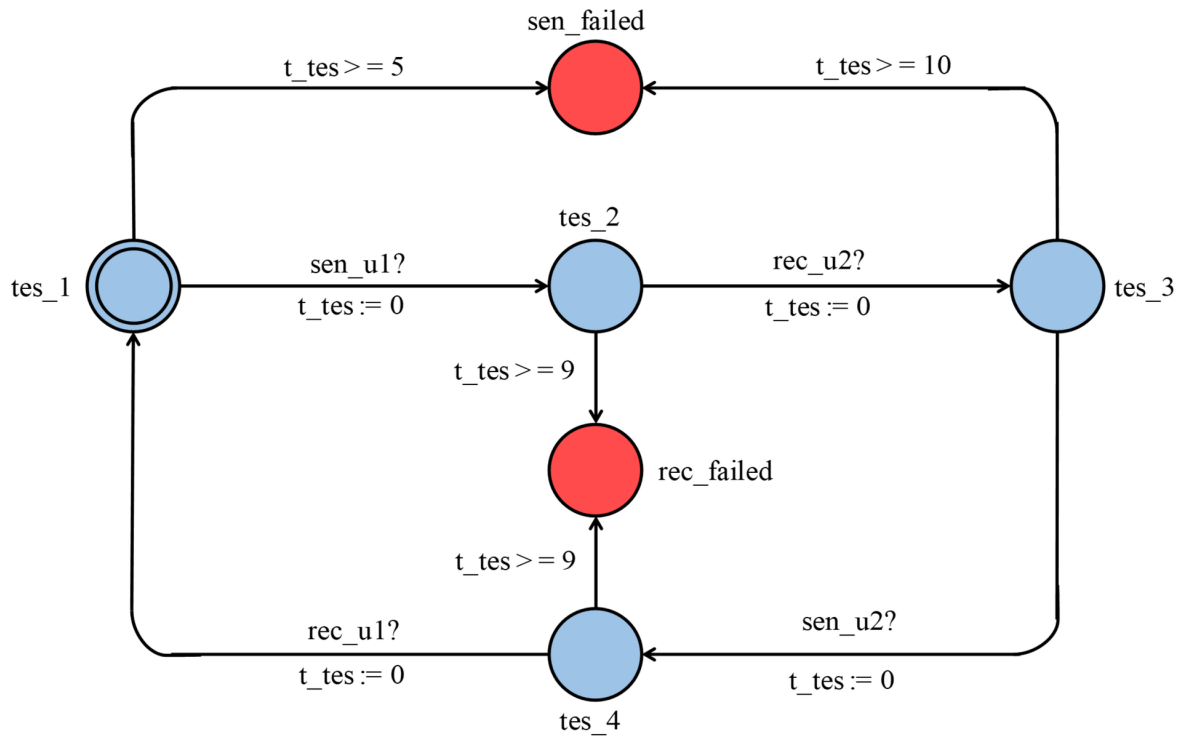


Figure 8. UPPAAL Model for Test Automaton



The verification of the above property guarantees that there would be no collision between the master and the slaves over the transmission channel. Next, we verify the bounded liveness property as:

$$\forall [] (T_Autom.sen_failed \vee T_Autom.rec_failed)$$

The verification of the above property certifies that the user-to-user communication delay is bounded. i.e., the system will transmit data within a time limit. Finally, we used the notion of deadlock present in UPPAAL to verify the deadlock freeness of the ZigBee protocol as:

$$\forall [] (\text{not deadlock})$$

Equivalently, the formal verification of this property also ensures the collision avoidance and bounded liveness properties, i.e., the system will never enter in the deadlock states namely `coll_state`, `sen_failed` and `rec_failed` of the master and test automata, respectively. The verification of all the above-mentioned properties ensures the true functional behavior of the smart grids' ZigBee-based routing protocol.

One of the distinguishing features of our formal analysis presented above is that all the state-space is exhaustively explored during the process of verifying the ZigBee protocol. Whereas, in the case of the simulation-based analysis, only some test vectors are explored, which compromise the accuracy of the analysis. Moreover, the verification is done automatically.

CONCLUSION

In this chapter, we presented the formal verification of the ZigBee-based routing protocol, which is widely used in HANs of smart grids to provide secure and reliable connection between the smart meter and home appliances. The UPPAAL model checker was used to model the master-slave architecture of the protocol using real-time automata. We also verified some vital properties of the ZigBee protocol, namely collision avoidance, bounded liveness and deadlock freeness properties, which ensure the correct functioning of the protocol. Our formal analysis of the ZigBee-based routing protocol using UPPAAL illustrates the advantages of using model checking for analyzing smart grid routing protocols.

FUTURE RESEARCH DIRECTIONS

In future, we plan to verify some other routing protocols for smart grid systems, such as wirelessHART (Song et al., 2008) and Z-Wave (Galeev, 2006) routing protocol for HAN. WirelessHART is a secured time-division multiple access (TDMA) based wireless mesh networking technology. Similarly, Z-Wave is used to integrate sensors and actuators and to perform the smart grid HAN services. The other research direction is to use our proposed approach for formally verifying the routing protocols for WAN and NAN for smart grids.

REFERENCES

- Baier, C., & Katoen, J. P. (2008). *Principles of Model Checking*. MIT Press.
- Ballarini, P., & Miller, A. (2006). *Model Checking Medium Access Control for Sensor Networks*. In *Leveraging Applications of Formal Methods, Verification and Validation*. IEEE.
- Behrmann, G., David, A., & Larsen, K. G. (2004) A Tutorial on UPPAAL. *Formal Methods for the Design of Real-time Systems*, (pp. 200–236). Springer Berlin Heidelberg.
- Bhargavan, K., Obradovic, D., & Gunter, C. A. (2002). Formal Verification of Standards for Distance Vector Routing Protocols. *Journal of the Association for Computing Machinery*, 49(4), 538–576. doi:10.1145/581771.581775
- Chen, Z., Zhang, D., Zhu, R., Ma, Y., Yin, R., & Xie, F. (2013). A Review of Automated Formal Verification of Ad-hoc Routing Protocols for Wireless Sensor Networks. *Sensor Letters*, 11(5), 752–764. doi:10.1166/2013.2653
- Clarke, E. M., Emerson, E. A., & Sistla, A. P. (1986). Automatic Verification of Finite-state Concurrent Systems using Temporal Logic Specifications. *ACM Transactions on Programming Languages and Systems*, 8(2), 244–263. doi:10.1145/5397.5399
- Clarke, E. M., Grumberg, O., & Peled, D. (1999). *Model Checking*. MIT press.
- Farhangi, H. (2010). The Path of the Smart Grid. *Power and Energy Magazine, IEEE*, 8(1), 18–28. doi:10.1109/MPE.2009.934876
- Fehnker, A., Fruth, M., & McIver, A. (2009). Graphical Modelling for Simulation and Formal Analysis of Wireless Network Protocols. *Methods. Models and Tools for Fault Tolerance*, 5454, 1–24. doi:10.1007/978-3-642-00867-2_1
- Fehnker, A., Glabbeek, R. V., Hofner, P., McIver, A., Portmann, M., & Tan, W. L. (2012) Automated Analysis of AODV Using UPPAAL. In *Tools and Algorithms for the Construction and Analysis of Systems*, (pp. 173–187). Springer.
- Fehnker, A., Hoesel, L. V., & Mader, A. (2007) Modelling and Verification of the LMAC Protocol for Wireless Sensor Networks. In *Integrated Formal Methods*, (pp. 253–272). Springer.
- Galeev, M. T. (2006). Catching the Z-Wave. *Embedded Systems Design*, 19(10), 28.
- Gao, J., Xiao, Y., Liu, J., Liang, W., & Chen, C. L. P. (2012). A Survey of Communication/Networking in Smart Grids. *Future Generation Computer Systems*, 28(2), 391–404. doi:10.1016/j.future.2011.04.014
- Gomez, C., & Paradells, J. (2010). Wireless Home Automation Networks: A Survey of Architectures and Technologies. *IEEE Communications Magazine*, 48(6), 92–101. doi:10.1109/MCOM.2010.5473869
- Gradara, S., Santone, A., & Villani, M. L. (2007). Formal Verification of Concurrent Systems via Directed Model Checking. *Electronic Notes in Theoretical Computer Science*, 185, 93–105. doi:10.1016/j.entcs.2007.05.031
- Hafeez, A., Kandil, N. H., Al-Omar, B., Landolsi, T., & Al-Ali, A. (2014). Smart Home Area Networks Protocols within the Smart Grid Context. *Journal of Communication*, 9(9), 665–671. doi:10.12720/jcm.9.9.665-671

- Harrison, J. (2009). *Handbook of Practical Logic and Automated Reasoning*. Cambridge University Press. doi:10.1017/CBO9780511576430
- Hasan, O., & Tahar, S. (2015). Formal Verification Methods. In *Encyclopedia of Information Science and Technology*. IGI Global.
- Hofner, P., & McIver, A. (2013). *Statistical Model Checking of Wireless Mesh Routing Protocols*. In *NASA Formal Methods*. Springer.
- Huq, M. Z., & Islam, S. (2010). Home Area Network Technology Assessment for Demand Response in Smart Grid Environment. In *Australasian Universities Power Engineering Conference*, (pp. 1–6). IEEE.
- Islam, S. M., Sqalli, M. H., & Khan, S. (2006). Modeling and Formal Verification of DHCP Using SPIN. *International Journal of Computer Science and Applications*, 3(2), 145–159.
- Kinney, P. (2003). Zigbee Technology: Wireless Control that Simply Works. *Communications Design*, 2, 1–7.
- Kwiatkowska, M., Norman, G., & Sproston, J. (2002) Probabilistic Model Checking of the IEEE 802.11 Wireless Local Area Network Protocol. In *Joint International Workshop on Process Algebra and Probabilistic Methods, Performance Modeling and Verification*, (pp. 169–187). Springer. 10.1007/3-540-45605-8_11
- Lundgren, H. (2002) *Implementation and Real-World Evaluation of Routing Protocols for Wireless Ad Hoc Networks*. Academic Press.
- Malik, S., Berthier, R., Bobba, R. B., Campbell, R. H., & Sanders, W. H. (2013). *Formal Design of Communication Checkers for ICCP using UPPAAL*. In *Smart Grid Communications*. IEEE.
- McCanne, S., Floyd, S., Fall, K., & Varadhan, K. (1997). *Network Simulator NS-2*. Academic Press.
- Orgun, M. A., & Ma, W. (1994). An Overview of Temporal and Modal Logic Programming. In *Temporal Logic*, (pp. 445–479). Springer.
- Pnueli, A. (1977). *The Temporal Logic of Programs*. In *Foundations of Computer Science*. IEEE.
- Qadir, J., & Hasan, O. (2015). Applying Formal Methods to Networking: Theory, Techniques, and Applications. *IEEE Communications Surveys and Tutorials*, 17(1), 256–291. doi:10.1109/COMST.2014.2345792
- Rashid, A., Hasan, O., & Saghar, K. (2015). *Formal Analysis of a ZigBee-based Routing Protocol for Smart Grids using UPPAAL*. In *High-Capacity Optical Networks and Enabling/Emerging Technologies*. IEEE.
- Renesse, R. D., & Aghvami, H. (2004) Formal Verification of Ad-Hoc Routing Protocols using SPIN Model Checker. *Mediterranean Electrotechnical Conference*, 3, 1177–1182. 10.1109/MELCON.2004.1348275
- Rossi, G. L. D., Gallina, L., & Rossi, S. (2013) Performance Analysis and Formal Verification of Cognitive Wireless Networks. In *European Workshop on Performance Engineering*, (pp. 236–250). Springer. 10.1007/978-3-642-40725-3_18
- Saputro, N., Akkaya, K., & Uludag, S. (2012). A Survey of Routing Protocols for Smart Grid Communications. *Computer Networks*, 56(11), 2742–2771. doi:10.1016/j.comnet.2012.03.027

Somappa, A. A. K., Prinz, A., & Kristensen, L. M. (2015). *Model-Based Verification of the DMAMAC Protocol for Real-time Process Control*. Verification and Evaluation of Computer and Communication Systems.

Song, J., Han, S., Mok, A., Chen, D., Lucas, M., Nixon, M., & Pratt, W. (2008). *WirelessHART: Applying Wireless Technology in Real-time Industrial Process Control*. In *Real-Time and Embedded Technology and Applications*. IEEE.

Tschirner, S., Xuedong, L., & Yi, W. (2008). *Model-based Validation of QoS Properties of Biomedical Sensor Networks*. In *Embedded Software*. ACM.

Yi, W., Pettersson, P., & Daniels, M. (1994). *Automatic Verification of Real-Time Communicating Systems by Constraint-Solving*. Formal Description Techniques.

ADDITIONAL READING

Boca, P., Bowen, J. P., & Siddiqi, J. (2009). *Formal methods: State of the Art and New Directions*. Springer Science & Business Media.

Gordon, M. J., & Melham, T. F. (1993). *Introduction to HOL A Theorem Proving Environment for Higher-order Logic*. Cambridge University Press.

Hall, A. (2007). Realizing the Benefits of Formal Methods. *Journal of Universal Computer Science*, 13(5), 669–678.

Hossain, E., Han, Z., & Poor, H. V. (2012). *Smart Grid Communications and Networking*. Cambridge University Press. doi:10.1017/CBO9781139013468

Kuzlu, M., Pipattanasomporn, M., & Rahman, S. (2014). Communication Network Requirements for Major Smart Grid Applications in HAN, NAN and WAN. *Computer Networks*, 67, 74–88. doi:10.1016/j.comnet.2014.03.029

McMillan, K. L. (1993). *Symbolic Model Checking*. *Symbolic Model Checking*. Springer. doi:10.1007/978-1-4615-3190-6

KEY TERMS AND DEFINITIONS

Formal Methods: Formal methods are the computer-based techniques that involve the development of the mathematical model of the given system based on some logic and verification of its intended properties using deductive reasoning.

Home Area Network (HAN): HAN is located in the customer premises and provides a connection between the smart meter and the home appliances.

Model Checking: Model checking is a formal verification method that involves the state-space based modeling of the given system and rigorously verifying its properties specified in an appropriate logic.

Smart Grids: Smart grids are a digital upgradation of the traditional grids providing reliable, secure, safe and efficient power distribution and management.

Timed Automata: Timed automata are the finite automata extended with a set of clock variables and are used for modeling the timing aspects of the system.

UPPAAL Model Checker: UPPAAL model checker involves modeling the behavior of the given system as a network of timed automata and is extensively used for formally analyzing the real-time systems.

Zigbee Routing Protocol: ZigBee routing protocol is commonly used in wireless HAN and provides a reliable, secure and uninterrupted connection between various components of HAN.

QoS Provisioning for Multicast Routing and Channel Assignment in Wireless Mesh Network

Abira Banik

 <https://orcid.org/0000-0001-8219-3026>

Tripura University, India

Abhishek Majumder

Tripura University, India

INTRODUCTION

Wireless mesh networks (WMN) is defined by a group of nodes which uses a wireless mode of channel and form a mesh topology, where each individual node is connected to the other with a dedicated line of communication. The few distinguished features of WMN's are self-organization, spatial reuse, and fault tolerance. The ad-hoc connectivity enables the WMNs to be of low cost, ease of maintenance, reliability and robustness [Akyildiz et al., 2005]. The scope of development and research in the fields of wireless mesh network is wide open and the focus has been also widened from only effective channel assignment to channel assignment and multicast routing. An effective channel assignment algorithm in WMN maximizes the utilization of network bandwidth. The efficiency of the network performance is generally termed as quality of service (QoS). The main term QoS defines the performance level of a service by the network to the end users. It generally is associated with the behaviour of wireless node or service, throughput sensitiveness and delay sensitiveness in multimedia applications. The QoS provisioning can be simply stated as satisfying a set of network related parameters like jitter, battery life, latency, packet loss ratio and bandwidth. The QoS imparting parameters depends upon the scenario in which it is implemented. In the context of channel assignment the QoS parameters can be signal power, bandwidth, power, etc. on the other hand the QoS parameters for multicast routing are bandwidth, delay, jitter etc. As for voice communication the parameters are battery life, for video communications buffer space, jitter delay, packet loss ratio, for military applications the parameters are strict security parameters. The various real time applications of the day to day life use multicasting such as audio/video conferencing, online gaming, remote learning, webcasting, distance learning, distribution of financial data, billing records, software, newspapers and lots more. WMNs are very much suitable for supporting such type of multicasting application. But provisioning of QoS in multicast routing as well as efficient assignment of channels is a challenging issue. For solving this problem, many techniques have been developed. This chapter presents a detailed study of the QoS provisioning in the channel assignment, multicast routing and both in a network scenario. It also classifies the techniques based on whether it is applied during channel assignment, multicast routing or during both phases. Comparison of the techniques is also presented in a tabular format.

DOI: 10.4018/978-1-7998-3479-3.ch070

BACKGROUND AND CLASSIFICATION

The incorporation of the QoS provisioning in the field of channel assignment and multicast routing techniques is already implemented by a number of techniques. Based on the point where the techniques are applied, QoS provisioning techniques in WMN can be classified in three categories namely, QoS provisioning in channel assignment, QoS provisioning in multicast routing and QoS provisioning jointly in channel assignment and multicast routing. A brief discussion on the already implemented techniques for serving QoS in the channel assignment phase, in multicast routing phase and channel assignment and multicast routing phase has been presented in the joint section.

QoS Provision in Multicast Routing Algorithm

In this section a brief discussion on the techniques for QoS provisioning in multicast routing algorithm has been presented.

Gateway Cluster Based Load Balancing Multicast algorithm

Gateway cluster based Load Balancing Multicast algorithm (GLBM) [Zhao et al., 2010] provides load balancing aware multicast routing algorithm which offers QoS as well as superior performance in terms of delay, jitter and throughput. In this algorithm, the main importance has been laid on the multicast routing communication pattern, since it has a high demand on network bandwidth and also an impact on algorithm execution time. It defines the WMN network in three distinct layers – Internet Gateway layer (IGW), Mesh Router layer (MR) and Mesh Client layer. The algorithm works on the IGW layer and MR layer. As stated above, the algorithm throws light in the concept of load balancing, since the mesh routers is responsible for maintaining the traffic load of the mesh clients. Likewise the gateways maintain the traffic load of the mesh routers and internet. It works as an admin in accessing the backbone internet and maintaining the optimal load balancing. Basically for any multicast routing scheme the lowest hop count route is selected as the best route, but in this scheme the best route is selected in terms of load balancing. The steps in the algorithm are:

- i. Initiating a multicast session – The source node in the multicast group initiates a multicast session to the internet, and sends a Multicast Sender Request (MSR) to its gateway to become a multicast source node. The gateway on receiving the MSR message periodically broadcasts the multicast Hello Message (HM) along with the other routing details like sequence number to other gateway nodes in the network. The gateway node on receiving the HM message, records the backward route to the source node in its multicast routing table and deletes the old entry.
- ii. Receiver joins the multicast session – The receiver which aims at joining in the multicast group sends the Multicast Route Request (MRQ - J) with Join flag and destination IP address to its gateway by broadcasting the request to all the neighbouring nodes at one hop distance. The node receiving this request (MRQ - J) adds the network Interface Queue (IFQ) length to the IFQ aggregation field. The path with the lowest IFQ aggregation is chosen as the multicast path from the gateway to the receiver. The IFQ length determines how busy the network interface of the node is. The gateway checks the multicast routing table. If the IP address is already available in the table then it sends a Multicast Route Reply (MRP) to the request receiver otherwise it sends Multicast Source Reply with Error flag (MSQ-E). The MSQ-E message is to inform the non-existence of the requested multicast session.

- iii. Maintaining multicast session – Whenever the multicast receiver is willing to leave the multicast group, it sends a Multicast Route Request with Leave flag (MRQ-L) message using default multicast route. The gateway node on receiving the MRQ-L deletes the entry of the particular node from its multicast routing table and notifies the multicast source using the Multicast Source Reply with Leave flag (MSQ-L). The multicast source on receiving the MSQ-L message deletes the entry for the receiver node from the multicast routing table and thereby stops sending any further multicast packets to this node.

The algorithm also incorporates the Error repair mechanism. If any receiver r does not receive packets from the multicast source s for a certain period of time, t , the node assumes that the link has broken or s has stopped multicast process. The receiver node r then sends out Multicast Route Request with Repair flag (MRQ-R) to its gateway node. The gateway on receiving the MRQ-R message, checks whether the source node is still forwarding multicast packets or not. If so, it chooses the best load route, updates the route and replies with MRP to the source s , otherwise sends MSQ-E to receiver notifying that multicast session no longer exists.

Hybrid QoS Multicast Framework-Based Protocol

The protocol [Pourfarkhar et al., 2010] combines proactive and reactive multicast routing in order to eliminate unnecessary delays as well as minimize control overhead in setting up routes in backbone mesh network. The main idea is to achieve gateway load balancing in multicast communication of WMN. Cerebral Model Articulation Controller (CMAC) [Albus et al., 1975] neural network model is used for load balancing and traffic oscillation in gateways during QoS multicast routing in WMN. The network model assumed in this protocol has more than one access point. Each multicast session is assigned a multicast group ID i.e. a multicast IP address in order to join or leave the multicast group. Route optimization and load balancing of network are the aims of the proposed protocol. The framework requires neither any knowledge of multicast member's location nor any modification to the existing Internet multicast routing protocol or IP multicast. The paths between the WMN and internet are chosen which guaranty least overloaded AP. The chosen path is backed with minimal delay and congestion. It elevates the packet delivery reliability. The bottleneck nodes can be avoided by distributing the traffic properly and consequently congestion can be controlled. The disconnection emphasis of the APs to the other nodes is higher because of the higher traffic load, so the nodes closer to the APs have more disconnection emphasis than the nodes situated at a distant.

For each node with a distance of h hops to its APs in the network Emphasis Coefficient (E_h) is calculated.

$$E_h = \begin{cases} 3^{h-m} & \text{if } (h < m) \\ 1, & \text{else} \end{cases} \quad (1)$$

If $h = 0$ then $E_h = 3^m$.

The congestion measurement for each node or simply Route Disconnection Expectancy (RDE) between the nodes and its AP with a distance of h hops is calculated as,

$$RDE_h = \sum_{i=0}^h (E_i \times NDP_i) = E_h \times NDP_h + RDE_{h-1} \quad (2)$$

where, NDP denotes Node Disconnection Probability

The path with minimum RDE is the least congested path to the AP with minimum workload. For fast recovery process and to have a load balanced network, each node records characters of another suitable AP as a reserved AP in a separate entry in its routing table along with the main AP. At first AP's are assigned to the mesh routers. The process is stated below,

- i. Each AP in the node announces its presence and status by broadcasting a status-notification messages to all nodes in the mesh at the configuration or reconfiguration time or when its workload increases and its probability of disconnection is high.
- ii. Each node after receiving status-notification messages calculates NDP and adds its RDE in the message packet. Then it records information of the two AP's, main AP and reserved AP in its AP routing table.
- iii. If node receives multiple messages from same AP then one of the neighbours can be chosen as the next node on the path to the AP based on shortest delay (first received message). If the node receives messages from different APs then main AP and reserved AP is selected based on least hop count or least RDE path. If both the nodes has routes with same RDE then choose AP with shortest path or route with minimum hop count is chosen.
- iv. If the selected AP is overloaded the message will be dropped. If a node receives overloaded message from the main AP, then it uses its reserved AP despite using its main AP, until traffic load on its main AP decreases.
- v. Each node after receiving new messages, updates the AP routing table for the main AP and reserved AP entries.

The entire process of assignment of mesh routers to the AP aims at a load balanced network with fast speed recovery, minimum delay routes and routes with minimum congestion.

The Hybrid QoS Multi-cast Routing Framework-Based Protocol for Wireless Mesh Network (HMFPM) supports dynamicity and scalability for wireless networks. The HMFPM works on the following three cases-

- Case I: The sender and receiver is in the same WMN.
 - Case II: The sender is in Internet and a subset of receiver in the WMN.
 - Case III: The sender is in WMN and a subset of receiver in the Internet.
- Case I

1. Before initiating the session, source S unicasts registration request message to its AP through the best route to register the information regarding group G and source S. The AP then broadcasts the information to the other AP in the network.
2. The source S floods the Receiver Discovery (RDS) data packet and buffers subsequent multicast data packet for a short time. All the nodes which receive the RDS packet, forward it and record the previous hop address. When a receiver R receives this packet, it unicasts a Receiver Join (RJN) packet back to S. The nodes which receive and forward the packet towards S, set up themselves as forwarder for this source and group. If a node receiving RJN has a routing table entry for this group and a source indicates that it is a forwarder, then the node will not forward the RJN to the source rather discards it.
3. When a receiver wants to join a group G, it sends a request information message to its AP to ask for the source of group G. The AP then replies with the group information (G, S). If the group is

not a source specific multicast group and there are many source for group G then the AP replies with the group information (G, Init_Val). On receiving the information of the group, if R has a path between it and source S, it unicasts a Join Request (JRQ) packet to S else, floods the network by broadcasting JRQ packet.

4. The intermediary nodes (M) process the received JRQ under three cases –
 - a. If M is neither a member of multicast group nor a forwarder node for source S in G, but has a path to the source S, then it must unicast the JRQ to the source.
 - b. If M is not a forwarder node and has no path to the source S in G, then it must broadcast the JRQ packet in the network.
 - c. If M is a forwarder node in the multicast group G and has received a unicast JRQ, it will not forward the JRQ to source. It creates a Join Reply (JRP) to itself and unicasts it to the neighbour node towards the JRQ sender. If it has received a broadcast JRQ, M only unicasts it to the S.
5. The source S on receiving the JRQ unicasts back a JRP to R.
6. The multicast source S will send the data packets to the forwarding nodes in the multicast group G. Thus the packets are forwarded to the receivers along the shortest delay path through the multicast tree of the multicast group.

Case II

If a receiver R in the WMN wants to join the multicast group G. It sends a request to its AP to ask for the source S of the group. The AP has not yet received any registration message from the sender, it learns that the sender is located outside the WMN. So, it acts as a proxy in the WMN for the Internet Sender. The source AP replies to R in the WMN. All the nodes that receive and forward packet towards R create a forwarding entry in their routing table for source AP and group G and sets the forwarder flag to True. The AP joins the multicast group G in the Internet using the existing Internet multicast routing protocol. On receiving the data packets from the internet sender the AP forwards the packets to the receivers in the WMN using the multicasting tree constructed by the HMFPM protocol.

Case III

The source S sends a message to its main AP to register the group information (G, S) before initiating any multicast session. The AP starts the multicast session in the Internet using the existing Internet multicast routing protocol as if it is the source of group G. After the receiver R in the internet joins the multicast session the AP uses HMFPM protocol to find out the route from source S to itself. The AP on receiving the packet from the sender multicasts it to the receivers in the Internet. The AP joins or leaves the multicast tree in the WMN depending upon the presence of the receivers in the Internet using the HMFPM protocol.

The two main advantages of the HMFPM are:

- Traffic migration – If one gateway stops functioning, the traffic can be routed and handled by another gateway in the WMN.
- Load balancing – Data traffic is distributed among the gateways based upon the associated metrics.

QoS Multicast Routing Algorithm

This technique [Ke et al., 2007] is a multi-constrained QoS based multicast routing algorithm. The network model is presented as a graph G , comprising a set of nodes, V , and a set of duplex links, E , between the node pairs. The goal of this technique is to find an optimal multicast route satisfying the necessary QoS parameters. The optimality conditions considered to select the paths are:

- i. Guaranteed bandwidth requirement,
- ii. End to end delay requirement satisfied,
- iii. Minimal cost.

In this scheme Raidl technique [Raidl et al, 2003] is used to represent the multicast tree. It is spanning tree used for evolutionary search, which compares the multicasting tree with some typical previous coding of spanning tree.

Each node $v \in V$ has a transmission range r . An edge $e(v_1, v_2) \in E$ between two nodes $v_1, v_2 \in V$ exists if distance between them is less than or equal to r .

The QoS parameters are defined as-

- (1) $\text{bandwidth}(t) = \min\{\text{bandwidth}(e(v_i, v_j)), e(v_i, v_j) \in t\}.$
- (2) $\text{Delay}(t) = \sum_{e(v_i, v_j) \in t} \text{delay}(e(v_i, v_j)).$
- (3) $\text{Cost}(t) = \sum_{e(v_i, v_j) \in t} \text{cost}(e(v_i, v_j))$

Here, t is a multicast tree in a set T . T is a set of possible multicast tree in G which connects the source S to every receiver $r_i \in R$. Genetic algorithm is a guided random search and optimization technique based on the basic principles of evolution i.e. survival of the fittest and inheritance. The three important aspects of the genetic algorithm i.e. selection, crossover and mutation are adopted to create individuals of the next generation from the current population.

The multicast tree is created randomly by RandomWalk algorithm. At every step, the algorithm moves over a randomly chosen adjacent edge to one of its neighbours. When all the destinations are found the search stops. The process of selection, crossover and mutation is discussed as follows:

- Selection: For the selection process Roulette wheel selection technique is adopted. This scheme considers the widened number and diverse solutions. The crossover probability P_c is assumed as 0.70 and the cataclysmic probability P_m is assumed as 0.35. It basically employs a heterogeneous recombination and cataclysmic mutation algorithm (CHC) inspired adaptive search [Eshelman et al, 1991] cross-generation selection competition mechanism which guarantees to present the best multicast tree selection. CHC is an evolutionary algorithm with binary coding which tries to avoid premature convergence by getting suitable balance between the ability to explore the search space for diversity and the ability to explore the local properties of the search for an appropriate selection [Luzon et al, 2004].
- Crossover: It is undergone to provide good quality solutions through inheritance from the parent solutions. Two trees are selected from the solution set and RandomWalkMulticastTree(RandomWalkMT) is used to get two off springs.

- **Mutation:** In this sub process a node is chosen randomly from the multicast tree. The edges are dropped which are connected to a node from the neighbouring node. The intersection of adjacencies of the pre-node and next-node of a node are drawn. If the adjacency of both the nodes intersect, then drop the node, else choose a node randomly from the intersection and add the two edges to the multicast tree.

The routing process is inspired by genetic algorithm where the size of the population and number of generation in the population are pre-determined. The initial set of multicast are solutions obtained from the RandomWalkMT. The multicast trees are encoded with the edge-set and different chromosomes are produced. The fitness value of the individuals of the current population is checked by evaluating the fitness function. The individuals of the current population and next population are mixed. Crossover operation is performed followed by mutation over the individuals of the mixed population. The outcomes of the previous steps are evaluated using the fitness function. Considering fitness functions outcome, the best individuals are selected from the mixed population and the current population is generated after the genetic operations. If the current and next population contains the same individuals then cataclysmic mutation is performed to increase the diversity of the solutions. The fitness function comprises of bandwidth parameter and delay cost of the multicast tree.

QoS Multicast Routing Framework

This algorithm [Shittu et al, 2008] addresses the problem of QoS provisioning in multicast routing scenario in WMN. It assumes a working QoS MAC layer protocol and proposes a framework which ensures QoS multicasting at network layer. It uses a combination of proactive or mesh and reactive or tree multicast routing based on backbone mesh routers and client nodes. This approach caters the QoS parameters like delay and congestion of data packets between backbone mesh and client STAs devices. The algorithm takes into consideration a network architecture comprising gateway which is connected to the internet in wired mode. The gateways are multicast sources and are connected to the mesh point in wireless mode. The mesh points act as mesh access points to a group a multicast receivers. The algorithm uses a routing protocol which states that a multicast path is always available before QoS request is done between the backbone mesh routers. This process eliminates the delay in finding the path between the routers. It also in turn minimizes channel access contention problems. The path establishment between the mesh access point (MAP) and end client station (STAs) in multicast group is done after the request is made. The MAP keeps an up-to-date record of link's conditions in the backbone mesh. This process minimizes the routing overheads.

QoS Provision in Channel Assignment Algorithm

In this section a brief discussion on the techniques for QoS provision in channel assignment algorithm has been presented.

Escalating QoS With Channel Assignment and Traffic Scheduling in Wireless Mesh Network

This technique [Salvakumar et al, 2018] uses tabu search heuristic approach for channel assignment in mesh network. The intention is to execute an intelligent exploration of search space that would ultimately get fascinated into local optima.

The two most distinguished features of the Tabu Search method are-

- Adaptive memory that takes intelligent decisions.
- Responsive exploration based on the unsurpassed solution in elongated time run in search space.

For a given explanation s , the neighbourhood of s is denoted by $n(s)$, defined by a set of realistic solution accessible from s by applying.

$$n(s) = \{s' | s' = m(s), m \in M(s), s' \in S\} \quad (3)$$

where, S = Resolution Space.

$M(s)$ = Set of association that can be generated from the solution s .

m = Movement to be applied to s .

The TabuSearchOptimization(TSO) technique is capable of holding memory of all the nodes in the network and assign channel which is interference free, at that particular time t . If there is any external interference in the network the TSO technique suggests the next node in the neighbour nodes for faster transmission of data. The most important feature of TSO technique is memory. The traffic scheduling in this technique is achieved in distributed manner, by sharing of common meta-data links. The traffic scheduling algorithm comprises two processes i.e. CoS quantization of traffic architecture and prioritization of connection request. The CoS quantization of traffic architecture sub process prioritize incoming traffic to guarantee Service Level Agreement. The incoming traffic is classified as:

- i. EF traffic as real traffic
- ii. AF traffic for non-real traffic
- iii. BF traffic for delay tolerant traffic.

The prioritization helps in dropping low priority traffic and identifying the delay tolerant traffic such as file transferring traffic, etc. The next sub process is the prioritization of connection request which is achieved by extracting two parameters from the connection requests- data rate consumption and the number of hops utilized in the network.

The parameters that are catered in the algorithm are hop count, tolerable delay, required data rate, and service priority. The hop count counts for the number of nodes to traverse before reaching the destinations. The tolerable delay is the delay that a connection request can tolerate. It is with reference with the QoS degree accepted at the destination. The minimum data rate required by the connection request throughout the Wireless Mesh Network. The service priority is the classification type of service request that a client is desirous to transmit or not.

Capacity and Interference Aware Link Scheduling with Channel Assignment

This algorithm [Kumar et al, 2011] finds out the traffic flow on each link and assigns the appropriate channel to link for minimizing the interference effect. The problem of channel assignment in this algorithm is formulated using linear programming along with the other related constraints and a new metric named link cost metric (LCM) which is an extension of the metric ETX (Expected Transmission count). In this approach, optimization of channel assignment and link scheduling is done. It also ensures minimum interference. The metric LCM captures the quality of the wireless link. The quality of the link is decided taking into consideration a number of constraints namely, capacity constraint, rate allocation per link, degree constraint and fairness constraint. The capacity constraint ensures that the total traffic passing through the router does not exceed the total capacity of all the channels. The rate allocation per link constraint ensures the minimum rate allocation for active link and maximum allowable time slot per link. The degree constraint checks whether a node is transmitting and receiving at the same time slot. The fairness constraint ensures fairness in traffic distribution. Given a WMN the main goal is to assign each communication link such that number of channels assigned to a particular link should not exceed the limit of radios in a node. The algorithm starts with a link assignment matrix (LAM) where all the group are sorted according to the value of link cost matrix (LCM). The link having the minimum value of LCM is assigned a channel. If there is an interference of links from different groups then assign the channel to the interfering channels. Else select the channel from the channel pool. Sort the interfering links thereby select the link having maximum interference. If there is any priority constraint then assign channel to the link having high priority to avoid further interference. This process is iterated for all the links in all groups of the network. The outcome of the algorithm is a channel assignment matrix (CAM) of the network. The QoS metrics served by this algorithm are – aggregate throughput, packet loss rate, packet delivery fraction (it is defined as ratio the number of packets received to the number of packets transmitted), average end-to-end delay and fairness index.

QoS Provisioning Jointly in Channel Assignment and Multicast Routing Algorithm

In this section a brief discussion on the techniques for QoS provisioning jointly in channel assignment and multicast routing algorithm has been presented.

Quality of Service Channel Assignment and Multicast Routing Algorithm

Quality of Service Channel Assignment and multicast Routing (Q-CAR) algorithm [Chakraborty et al., 2017] is an intelligent solution for joint QoS multicast routing and channel assignment in multichannel multi radio wireless mesh networks. The multicast routing mechanism in multichannel multi radio WMN's has been proved to be a NP-Hard problem. The multicast routing mechanism leads to construction of multicast tree which is a very important unit in multimedia and real time data transfer. It also plays an important role in achieving optimal performance in multi-channel multi radio WMN's scenario. The Q-CAR method solves the multicast tree construction and channel assignment problem conjointly by intelligent differential evolution. Differential evolution is a form of stochastic algorithm for optimization inspired by biological processes which learn from their surrounding environment. The channel assignment algorithm is a slightly modified form of the basic differential evolution approach. The intelligent multicast tree construction phase is based on the genetic algorithm approach. The aim of the scheme is to assign

channels with minimum channel conflict. The fitness of the outcomes of the genetic approach used in the channel assignment algorithm is evaluated using the Signal to noise ratio (SNR) as the fitness measuring parameter, which enhances the overall efficiency. The SNR is the ratio of received signal strength and background noise. The optimization model for channel assignment algorithm is:

$$\text{minimize } f(X_{i,G}) = \frac{1}{SW(T_r) + 1} \quad (4)$$

subject to,

$$\forall r_j \in R_{Tr} \left[\sum_{l_{ik} \in P(s,rj)} \delta_{l_{ik}} \leq \Delta \right]; \delta_{l_{ik}} \geq 0 \quad (5)$$

$$\forall r_j \in R_{Tr} \left[\sum_{l_{ik} \in P(s,rj)} \varepsilon_{l_{ik}} \leq \Delta \right]; \varepsilon_{l_{ik}} \geq 0 \quad (6)$$

Where $SW(T_r)$ is the total signal weight for a routing tree R_{Tr} .

$\delta_{l_{ij}}$ is the average delay experienced at link l_{ij} , $\varepsilon_{l_{ij}}$ is the average jitter experienced at link l_{ij} .

The goal of effectively providing QoS is achieved by subjecting the outcomes of the channel allocation algorithm to the delay and jitter upper bounds. Then the second phase of multicast tree construction is initiated using the outcomes of the previous step i.e. channel allocation. Because it has been seen in the other existing approaches that if the multicast routing outcomes are fed as the input to the channel assignment phase, it causes a major performance degradation. The interference disturbance of the nodes are sensed and the cost of the entire tree constructed is calculated to evaluate the fitness of the units of the genetic algorithm approach (here chromosome).

The optimization model for the multicast routing phase is,

$$\text{maximize } F(Ch_i) = \frac{1}{\beta ID_{Tr} + (1 - \beta) Cost_{Tr}} \quad (7)$$

subject to

$$ID_{Tr} \geq 0, Cost_{Tr} \geq 0; \forall T_r, \beta \in [0, 1]$$

$$\forall r_j \in R_{Tr} \left[\sum_{l_{ik} \in P(s,rj)} \delta_{l_{ik}} \leq \Delta \right]; \delta_{l_{ik}} \geq 0 \quad (8)$$

$$\forall r_j \in R_{Tr} \left[\sum_{l_{ik} \in P(s, r_j)} \varepsilon_{l_{ik}} \leq \Delta \right]; \varepsilon_{l_{ik}} \geq 0 \quad (9)$$

Where, ID_{Tr} is the total interference disturbance on tree T_r . $Cost_{Tr}$ is the tree cost.

The multicast treeconstruction phase is a reactive process, i.e. a multicast tree is built whenever a request for themulticast communication arises. In this method at the generation of the initial popula- tion, thefirst four participating chromosomes are generated using the Prim's algorithm, BFS, DFS and Dijkstra shortest path algorithm in order to speed up the convergence in generating potentialmulticast routing tree. The above mentioned four algorithms generates four candidate solutionshaving minimum interference impact and tree cost, thereby helping in generating better solutionsrather than using random combinations [Chakraborty et al., 2015]. The units of the initial population are succeeded bythe crossover and mutation process thereby generating the entire population of parent population. The most important aspect of the genetic algorithm is the selection of the fittest units to passthem to the next generation depending upon the fitness function. Linear ranking selection schemeis used for selection.

Intelligent Quality of Service Multicast Routing and Channel Assignment Algorithm

The intelligent Quality of service multicast routing and channel Assignment (i-QCA) algorithm [Chakraborty et al., 2015] solvesthe multicast tree construction and channel assignment problem by involving intelligentcomputational methods. This technique involves genetic algorithm for channel as- signment which caters the co-channel conflict for the tree and delay, jitter and minimal interference low cost multicast tree. It basically uses the differential evolution approach in the construction ofthe multi- cast tree and genetic algorithms inspired channel assignment algorithm. The algorithm builds a low cost multicast tree and the QoS determining parameters are delay and jitter bounded minimal interference. It also tries to minimize the total co-channel interference andthe number of the scarce radio interfaces. Here other than the co channel interference the scheme considers radio interfacesbecause in a multicast communication the source has a single radio interface thatsends the multicast packets and the intended receivers deploy one radio interface to receive thepackets. The intermediary nodes perform the both, send and receive. Therefore, the scheme tries to reduce the number radio interface thepacket has to go through before reaching the destination nodes. The optimization model for the channel assignment algorithm is,

$$\text{Maximize } f(Ch_i) = \frac{1}{cc(T_r) + 1} \quad (10)$$

Subject to,

$$cc(T_r) \geq 0; \forall T_r$$

$$\forall r_j \in R_{Tr} \left[\sum_{l_{i \in P(s, r_j)}} \delta_i \leq \Delta \right]; \delta_i \geq 0 \quad (11)$$

$$\forall r_j \in R_{Tr} \left[\sum_{l \in P(s, r_j)} \leq \Delta \right];_i \geq 0 \quad (12)$$

The tree cost is calculated using number of deployed radio interfaces. The fitness function evaluates the quality of a multicast tree in terms of total interference impact and radio interface cost of the multicast routing unit. It is a method which solves the channel assignment and multicast routing in polynomial time.

Quality of Service Multicast Routing and Channel Assignment

The Quality of Service Multicast Routing and Channel Assignment (QoS-MRCA) scheme [Cheng et al., 2011] is concerned with the quality of service multicast routing and channel assignment, thereby giving joint solution for the multicast routing problem. The method includes three algorithms namely genetic algorithm, simple annealing and tabu search to search minimum interference multicast trees which also satisfies the end-to-end delay constraint and optimizes the usage of the scarce radio network resource in the WMN. The above mentioned method aims to build a delay bounded multicast tree and also to minimize total channel conflict. The QoS is provided in terms of both the total channel conflict and the tree cost. The multicast tree cost is used to evaluate the utilization of the network resource as well as treated as a QoS parameter in wireless networks where limited radios and channels are available. The fitness function evaluates the units of multicast routing in terms of total channel conflict and tree cost. This results in choosing a low cost multicast tree on which the minimum interference channel assignment is achieved. The main motive behind the implementation of the method is to improve performance in order to provide QoS to the users in various collaborative multimedia applications which require efficient information delivery service from a designated source to multiple receivers. The genetic algorithm is a stochastic meta-heuristic optimization method that involves methods and model from biological principles of Darwinian Theory of evolution and Mendelian principles of inheritance. The objective function involves minimization of the total channel conflict, in order to improve the system throughput. This scheme uses an effective channel assignment algorithm which gives least channel conflict on any multicast tree. For ease of presentation binary interference model is adopted where two links either totally interfere or don't interfere at all. In the multicast routing, a tree is evaluated with minimal tree cost, as the number of the radio interfaces involved in the multicast communications. In order to reduce the tree cost, the number of links used in the multicast routing tree and the end-to-end delay has been reduced. Regarding the fitness function, the primary criterion regarding solution quality is the total channel conflict and the subsidiary one is the tree cost. Thereby, among the candidate solutions with the same minimum channel conflict, and choosing the one with the lowest tree cost is chosen. The GA-based joint QoS multicast routing and channel assignment algorithm consists of sub processes like fitness function, selection scheme, crossover and mutation. The Simulated Annealing (SA) based joint QoS-MRCA algorithm consists of the sub processes like solution representation, neighbourhood structure, initial temperature, temperature decreasing, iterative length at each temperature and the termination rule. The TS operations consist of the sub processes like solution representation, initial solution, neighbourhood structure, fitness function, tabu move, tabu list, aspiration criterion and termination rule.

Utility-Based Multicast Routing and Channel Assignment

The Utility-based multicast routing and channel assignment (UBMR-CA) [Kumar et al., 2013] scheme provides an efficient and QoS aware routing and channel assignment service to the users also it incorporates features as low deployment and maintenance cost. In this scheme certain constraints are considered such as end to end delay, interference and available bandwidth which causes limitations to the overall QoS services availability to the end users. There are mainly two objectives of the scheme i.e. loop free routing and minimization of the total utility of a link in the network from a source to the destinations at any time. The term utility here is calculated using a utility weight metric (UVM) which is implemented for sending multicast messages from a single source to multiple destinations. The utility of the links are calculated using a utility weight metric (UVM) for sending and receiving messages through all links in the network. The channels for the links are assigned keeping in view the various constraints i.e. capacity, bandwidth and interference between different links. The utility of a link is the combination of its capacity, load and current interference with the neighbouring link. The utility of a link is subject to change with reference to changes in the capacity, link and interference with time. The UBMR-CA scheme constructs the utility based multicast tree, followed by the construction of the channel conflict graph and finally the utility based channel assignment. The UVM metric depends upon the number of interfaces, channel switching time, and number of available channels. The UVM is developed for multicasting and channel assignment according to the demand of the particular service required by the users. The metric has a great impact in aggregate throughput and delay which are the most critical for any application in WMN. The utility based channel assignment algorithm runs for initialization of an initial utility. The traffic patterns are divided into different groups, and depending upon it channels are allocated. In each time slot for every group the utility and channel assignment process are carried out. The channel assignment is not complete until utility of all members does not change in that group and in that time slot. The technique is also effective in achieving fairness between the various traffic flows, resulting in a good balance of the traffic load thereby improving the overall quality of the application and services provided by the WMNs.

Multicasting with Multiple Gateways and Partially Overlapped Channels

The Multicasting with multiple Gateways and Partially Overlapped Channels (MG-POC) framework [Shahmirzadi et al., 2016] is based on binary integer programming that provides framework for minimum interference in multicast communication. It provides a rational node selection to construct multicast tree that increases overall network performances. MG-POC efficiently constructs the paths between the source and the receivers by including minimal number of data forwarding nodes. It employs multiple gateways in order to reduce interference and usage of resources. Inclusion of the broadcast advantage of the wireless communication and usage of the partially overlapped channels in the channel assignment phase makes the scheme more advantageous. It solves the channel assignment and multicast tree construction problems simultaneously. MG-POC finds almost optimal solutions for large network problems in a reasonably short amount of time. It also supports multiple gateways and partially overlapped channels from the entire channel bandwidth. The framework also gives solutions that causes minimal network interference and enhances network throughput. MGPOC not only employs hop count as the routing metric in the multicast routing tree construction phase rather employs a rational node selection scheme. A Binary Integer Programming (BIP) is used to solve the problem, its efficiency and performance investigation. The rational node selection scheme constructs paths from a single source to the multiple receivers with a minimum

number of data relaying nodes. The MG-POC framework incorporates the Wireless Broadcast Advantage (WBA), which reduces interference and improves throughput. MG-POC also incorporates the multiple objective function with constraints as total number of forwarding nodes and the total interferences. A weakly decoupled approach is used to find a nearly optimal solution for fairly large network problems in a reasonably short amount of time. This technique uses physical interference model where communication between two nodes is successfully carried out only when the Signal to Interference and Noise Ratio (SINR) is above the permissible value or threshold. The multicast tree built in this scheme is a minimum cost tree which provides data transmission from Internet-connected gateways towards multicast receivers such that the multicast tree encompasses the set of all receivers related to the multicast communications. The multicast tree construction and channel assignment processes are solved sequentially where first the multicast tree is solved then the channel assignment process. In the multicast tree construction phase, a multicast tree is developed with minimum number of forwarding nodes. In the next phase the channel assignment is done where the total interference is taken as the objective function.

Multi Objective Optimization Approach for Joint Channel Assignment & Multicast

Routing

This scheme [Vaezpour et al, 2014] uses multi objective genetic algorithm to construct minimum cost multicast tree with minimal interference. It is achieved by optimizing the solution with constraints such as total channel conflict, tree cost, maximum end-to-end delay and average end-to-end delay. In other words, it produces a multicast tree with low interference, low cost and low delay. However the fourth objective, total channel conflict, is used in optimization of the channel assignment algorithm. For the overall improvement of the algorithm Evolutionary Multi Objective (EMO) algorithm namely, NSGA-II [Lobiyal et al, 2017] is used. NSGA-II is an optimization technique which can consider multiple optimization constraint at the same time and ranks the solution according to the non-dominance sorting and crowding distance. Since the multicasting data is transmitted the end-to-end delay, the network bandwidth usage, the delay jitter are specified in the multicast routing process. Reducing the overall cost reduces the resource usage as well as satisfies delay requirements thereby guarantying the data delivery with up to date data. Since data is transmitted along the multicast tree, the end to end delay, the network bandwidth usage, and the delay jitter are mainly specified by the multicast tree. The QoS is served as reducing the cost of the multicast tree helps in lowering resource usage. And the delay requirements of the multicast applications guarantees that the data delivered to the destination is up to date. Since the technique is NSGA-II inspired, the quality of the solutions against others is solved using Pareto ranking method. The Pareto dominance rule is used to sort the solutions and assigning them to the different fronts.

COMPARISON

In this section the QoS provisioning techniques have been compared. Table 1 shows the comparison of the schemes.

Table 1. Comparison of all the techniques, conditions applied and the QoS parameters served

Sl. no.	Algorithm	Conditions applied	QoS parameters
1.	GLBM[Zhao et al., 2010]	Multicast Routing	Packet loss, delay
2.	A Hybrid QoS multicast framework based protocol[Pourfarkhar et al., 2010]	Multicast Routing	Packet delivery ratio, packet delay, overhead.
3.	A QoS multicast routing algorithm[Ke et al., 2007]	Multicast Routing	Bandwidth requirements, end-to-end delay requirement and minimum cost.
4.	QoS Multicast Routing Framework [Shittu et al, 2008]	Multicast Routing	Delay, congestion overhead control.
5.	Escalating QoS with channel assignment and traffic scheduling [Salvakumar et al, 2018]	Channel Assignment	Throughput, energy, error rate.
6.	Capacity and interference aware link scheduling with channel assignment [Kumar et al, 2011]	Channel assignment	Aggregate throughput, packet loss rate, packet delivery, average end-to-end delay and fairness index.
7.	Q-CAR[Chakraborty et al., 2017]	Channel Assignment Multicast Routing	Interference disturbance, tree cost, SNR, delay, jitter.
8.	i-QCA[Chakraborty et al., 2015]	Channel Assignment Multicast Routing	Total co-channel conflict, tree cost, interference impact, radio interface cost.
9.	QoS-MRCA[Cheng et al., 2011]	Channel Assignment Multicast Routing	Channel conflict, tree cost.
10.	UBMR-CA[Kumar et al., 2013]	Channel Assignment Multicast Routing	End-to-end delay, interference, cost.
11.	MG-POC[Shahmirzadi et al., 2016]	Channel Assignment Multicast Routing	Interference, tree cost.
12.	Multi objective Optimization Approach for Joint Channel Assignment & Multicast Routing [Vaezpour et al, 2014]	Channel Assignment Multicast Routing	Total channel conflict, tree cost, end-to-end delay.

CONCLUSION

The main aim of this chapter is to study the QoS provisioning in wireless mesh networks when applied in different network phases, as channel assignment, multicast routing and in scenarios where channel assignment and multicast routing are jointly presented. The QoS is defined by the performance level of a service of a network with respect to some network parameters related to the service provided to the end users. QoS provisioning term is associated with the behaviour of wireless node or service of the network, throughput sensitiveness and delay sensitiveness in different applications of WMN. The inclination of research is towards QoS provisioning to the end users. As well as incorporating more independent parameters for overall improvement of the QoS of the network. The chapter presents different QoS scenarios in which QoS is expected by the end users of the network, requirements for providing QoS and challenges faced in the course of providing QoS. The QoS provisioning to the end users are further classified under three heads: QoS provision in multicast algorithm. QoS provision in channel assignment and QoS provision jointly in channel assignment and multicast routing. The existing techniques are discussed in brief under the three heads.

FUTURE RESEARCH DIRECTIONS

The algorithm, scenario in which it is applied and QoS parameter(s) served to the end users of the network is presented in tabular format. QoS provisioning and channel assignment considering more parameters such as traffic type is one of the future work directions. Moreover, such techniques for heterogeneous wireless networks also can be developed in future.

REFERENCES

- Akyildiz, I. F., Wang, X., & Wang, W. (2005). Wireless mesh networks: A survey. *Computer Networks*, 47(4), 445–487. doi:10.1016/j.comnet.2004.12.001
- Albus, J. S. (1975). A new approach to manipulator control: The cerebellar model articulation controller (CMAC). *Journal of Dynamic Systems, Measurement, and Control*, 97(3), 220–227. doi:10.1115/1.3426922
- Chakraborty, D. (2015). i-QCA: An intelligent framework for quality of service multicast routing in multichannel multiradio wireless mesh networks. *Ad Hoc Networks*, 33, 221–232. doi:10.1016/j.ad-hoc.2015.05.007
- Chakraborty, D., & Debbarma, K. (2017). Q-CAR: An intelligent solution for joint QoS multicast routing and channel assignment in multichannel multiradio wireless mesh networks. *Applied Intelligence*, 47(1), 13–27. doi:10.1007/10489-016-0871-2
- Chakraborty, D., Debbarma, M. K., & Roy, S. (2013). Qos provisioning in wmns: Challenges and a comparative study of efficient methodologies. *International Journal of Computers and Applications*, 65(3).
- Cheng, H., & Yang, S. (2011). Joint QoS multicast routing and channel assignment in multiradio multichannel wireless mesh networks using intelligent computational methods. *Applied Soft Computing*, 11(2), 1953–1964. doi:10.1016/j.asoc.2010.06.011
- Choi, S., & Park, J. (2016). Minimum interference channel assignment algorithm for multicast in a wireless mesh network. *Sensors (Basel)*, 16(12), 2056. doi:10.3390/16122056 PMID:27918438
- Eshelman, L. (1991). *The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination, Foundations of Genetic Algorithms*. Morgan Kaufmann.
- IEEE 802.11 Working Group. (1999). Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *ANSI/IEEE Std. 802.11*.
- Jahanshahi, M., & Barmi, A. T. (2014). Multicast routing protocols in wireless mesh networks: A survey. *Computing*, 96(11), 1029–1057. doi:10.1007/00607-014-0403-z
- Ke, Z., Li, L., Sun, Q., & Chen, N. (2007). A QoS multicast routing algorithm for wireless mesh networks. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. Eighth ACIS International Conference on* (pp. 835–840). 10.1109/SNPD.2007.186
- Kumar, N., Chilamkurti, N., & Lee, J. H. (2013). UBMR-CA: Utility-based multicast routing and channel assignment with varying traffic demands in multi-radio multi-channel wireless mesh networks. *Mathematical and Computer Modelling*, 57(11), 2750–2763. doi:10.1016/j.mcm.2011.12.009

- Kumar, N., Kumar, M., & Patel, R. B. (2011). Capacity and interference aware link scheduling with channel assignment in wireless mesh networks. *Journal of Network and Computer Applications*, 34(1), 30–38. doi:10.1016/j.jnca.2010.10.001
- Luzón, M. V., Barreiro, E., Yeguas, E., & Joan-Arinyo, R. (2004). GA and CHC. Two evolutionary algorithms to solve the root identification problem in geometric constraint solving. *International Conference on Computational Science*, 139–146.
- Pourfakhar, E., & Rahmani, A. M. (2010). A hybrid QoS multicast framework-based protocol for wireless mesh networks. *Computer Communications*, 33(17), 2079–2092. doi:10.1016/j.comcom.2010.07.026
- Raidl, G. R., & Julstrom, B. A. (2003, June). Edge sets: An effective evolutionary coding of spanning tree. *Evolutionary Computation*, 7(3), 225–239. doi:10.1109/TEVC.2002.807275
- Reddy, T. B., Karthigeyan, I., Manoj, B. S., & Murthy, C. S. R. (2006). Quality of service provisioning in ad hoc wireless networks: A survey of issues and solutions. *Ad Hoc Networks*, 4(1), 83–124. doi:10.1016/j.adhoc.2004.04.008
- Selvakumar, K., & Revathy, G. (2018). Escalating quality of services with channel assignment and traffic scheduling in wireless mesh networks. *Cluster Computing*, 1–7.
- Shahmirzadi, M. A., Dehghan, M., & Ghasemi, A. (2016). An optimization framework for multicasting in MCMR wireless mesh network with partially overlapping channels. *Wireless Networks*, 24(4), 1–19.
- Shittu, W. A., Hashim, A. H. A., Anwar, F., & Al-Khateeb, W. (2008). A proposed QoS multicast routing framework for next-generation wireless mesh network. *IJCSNS*, 8(9), 280.
- Vaezpour, E., & Dehghan, M. (2014). A multi-objective optimization approach for joint channel assignment and multicast routing in multi-radio multi-channel wireless mesh networks. *Wireless Personal Communications*, 77(2), 1055–1076. doi:10.1007/11277-013-1554-5
- Wang, J., & Shi, W. (2017). Joint multicast routing and channel assignment for multi-radio multi-channel wireless mesh networks with hybrid traffic. *Journal of Network and Computer Applications*, 80, 90–108. doi:10.1016/j.jnca.2016.11.022
- Zeng, G., Wang, B., Ding, Y., Xiao, L., & Mutka, M. (2007, October). Multicast algorithms for multi-channel wireless mesh networks. In *International Conference on Network Protocols*. IEEE. 10.1109/ICNP.2007.4375831
- Zhao, L., Al-Dubai, A. Y., & Min, G. (2010). GLBM: A new QoS aware multicast scheme for wireless mesh networks. *Journal of Systems and Software*, 83(8), 1318–1326. doi:10.1016/j.jss.2010.01.044

ADDITIONAL READING

- Al Islam, A. A., Islam, M. J., Nurain, N., & Raghunathan, V. (2016). Channel assignment techniques for Multiradio wireless mesh networks: A survey. *IEEE Communications Surveys and Tutorials*, 18(2), 988–1017. doi:10.1109/COMST.2015.2510164

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197. doi:10.1109/4235.996017

Draves, R., Padhye, J., & Zill, B. (2004, September). Routing in multi-radio, multi-hop wireless mesh networks. In *Proceedings of the 10th annual international conference on Mobile computing and networking* (pp. 114–128). ACM. 10.1145/1023720.1023732

Farmani, B., Jaseemuddin, M., & Batarfi, O. (2017, April). IMRP: Interference-aware multicast routing for Wireless Mesh Networks. In *Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on* (pp. 1–6). IEEE.

Kyasanur, P., & Vaidya, N. H. (2005, March). Routing and interface assignment in multi-channel multi-interface wireless networks. In *Wireless Communications and Networking Conference, 2005 IEEE* (Vol. 4, pp. 2051–2056). IEEE.

Lobiya, D. K., & Prasad, S. (2017). An elitist nondominated sorting genetic algorithm for QoS multicast routing in wireless networks. *Swarm and Evolutionary Computation*, 33, 85–92. doi:10.1016/j.swevo.2016.10.004

Ramachandran, K. N., Belding, E. M., Almeroth, K. C., & Buddhikot, M. M. (2006, April). Interference-aware channel assignment in multi-radio wireless mesh networks. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings* (pp. 1–12). IEEE. 10.1109/INFOCOM.2006.177

Saini, J. S., & Sohi, B. S. (2016). A Survey on Channel Assignment Techniques of Multi-Radio Multi-channel Wireless Mesh Network. *Indian Journal of Science and Technology*, 9(42).

Subramanian, A. P., & Buddhikot, M. M. (2006, September). Interference aware routing in multi-radio wireless mesh networks. In *Wireless Mesh Networks, 2006. WiMesh 2006. 2nd IEEE Workshop on* (pp. 55–63). IEEE. 10.1109/WIMESH.2006.288620

Yang, Y., Wang, J., & Kravets, R. (2005, September). Designing routing metrics for mesh networks. In *IEEE Workshop on Wireless Mesh Networks (WiMesh)* (pp. 1–9).

Yuan, J., Li, Z., Yu, W., & Li, B. (2006). A cross-layer optimization framework for multihop multicast in wireless mesh networks. *IEEE Journal on Selected Areas in Communications*, 24(11), 2092–2103. doi:10.1109/JSAC.2006.881617

KEY TERMS AND DEFINITIONS

Channel Assignment: Channel assignment is the process of assigning separate orthogonal channel or partially over-lapped orthogonal channels to all the nodes in the communication range. The main motive behind assignment of channels is to reduce the interference impact and proper delivery of packets on time. For the WMN in channel assignment phase, the entire range of channels in IEEE 802.11 (a, aa, b, g, n) standard is utilized for assigning channels in the links. By the term link, it means the links between the mesh nodes. Channel allocation plays the most crucial role in achieving better performance of the network.

Multicast Routing: Multicasting is a mechanism where messages are transferred over links in the WMN once and are duplicated at the branch or break points in order to successfully deliver them to the different destinations. Multicasting is a form of communication that helps in delivering information from a single source to multiple destinations nodes simultaneously in an efficient manner. The multicasting routing is the process of selecting a route from a single source to multiple destinations. The multicast routing or simply multicasting cannot be attained by utilizing the concept of combined unicast from source to individual destination or simplified reduced broadcast from the source to all the receivers (Zeng et al., 2007). Multicast can transmit the same data flow to multiple destinations without producing the duplicate flow and without wasting bandwidth of the network (Choi et al., 2016). The main unit of multicast routing is the multicast routing tree. The various applications of the multicast routing includes audio/video conferencing, online gaming, remote learning, webcasting, distance learning, distribution of financial data, billing records, software, newspaper and lots more. To achieve efficient multicast communications, reasonable routing paths should be selected which connects all multicast receivers with the source node and also assignment of proper channels to the links of the network should be done (Wang et al., 2017).

Evaluating the Effectiveness of Multi-Web Services in Load Balancing Cluster-Based Web Server

Abhijit Bora

 <https://orcid.org/0000-0002-7754-639X>

Gauhati University, India

Tulshi Bezboruah

 <https://orcid.org/0000-0002-9895-0973>

Gauhati University, India

INTRODUCTION

Web Service (WS) provides a flexibility platform for delivering computational task as service. Its architecture supports deployment of grid computing, Internet of Things (IoT) based applications and distributed computing in a flexible manner. The interoperable features of WS deployment support the homogeneous and heterogeneous communications among software modules.

It follows the principles of Service Oriented Architecture (SOA). The principles of SOA focus the methodology of service publishing, consuming and discovering over public registries. It supports service reuse and integration of additional service. The SOA is not limited to a technology. It can be implemented through the variety of programming languages, technologies and communication protocol. The power of Extensible Markup Language (XML) provides a framework in SOA for establishing communication among different business nodes. Recently, WS has become attractive paradigm for building the Software as a Service (SaaS). It is defined as SaaS for machine to machine interaction over different network topologies. The SOA offers the features of adaptability over different network connection and capabilities of terminals.

The principles of interoperability, reusability and modularity can be deployed through the methodology of WS. However, number of concerns and challenges are observed for the quality of the service (QoS). The QoS aspects can be observed from the perspective of service providers and consumers stakeholders. The service providers primarily concern about proper execution and implementation aspects of the service. Conversely, the service consumer primarily concerns about the quality assurance of the service. The service consumers also identifies whether the service is the right service for their business process or not. The service consumers usually identify the service over public service registries. The service can be executed and discovered at run time. Based on the functional parameter and business decision model, the consumer accesses the functionalities of the service. The service consumer is termed as WS client that can communicate the service provider and provide the additional service along with the existing service.

Many practitioners have been trying to deploy WS to deal with distributed business processes. Many prominent software industries are deploying their business services through the paradigm of WS architecture. Their primary goal is to deliver better and enhanced service to end users. With rapid growth of internet users, the consumer of WS is also increasing exponentially. As a service owner of WS, the

DOI: 10.4018/978-1-7998-3479-3.ch071

primary concern is to deliver the service to massive users even in high stress of usage. However, the load balancing cluster based web server can play a vital role for handling massive users of WS. The clustering methodology of web server using single or multiple machines can deliver reliable and efficient service to end users. With rapid growth of organizational infrastructure and independent service delivery modules, the deployment of multiple WS and assessing their effectiveness is becoming crucial among software practitioners and researchers. As such, the QoS against high load of users is becoming an important concern among research community. This work will introduce a novel QoS assessment framework for evaluating the effectiveness of multi WS deployment using load balancing cluster based web server.

BACKGROUND

WS provides the platform as framework for deploying software agents as service over network and enhances the functional execution of business process. It utilizes the different loosely or tightly coupled software services for business logics (BL). It provides conceptual as well as programmable interface for integration of modular oriented functional logic in a respective service (Matthew et al., 2005). The consumer of WS is not aware of internal data structure and the service platform that it is executing. The consumer simply utilizes the WS for the computational BL. The overall service that executes in server side and delivers service over network based protocol is termed as Service Oriented Computing (SOC). However, based on service delivery scenario and BL, the SOC based system may utilize the deployment of single or multiple WSs for service delivery. The SOC based system primarily utilizes SOAP, Web Service Description Language (WSDL), and Universal Description Discovery and Integration (UDDI). The software agents communicate through SOAP messaging that contains Extensible Markup Language (XML) nodes (SOAP, 2016). The publicity of the service is highlighted through WSDL file of the service. It establishes the connectivity and generates necessary server side files in client side (WSDL, 2016). The UDDI is the platform that showcases the required functional parameters of the service (Peiris et al., 2007). The block diagram of SOC based system and the different ways of WS communication are discussed elsewhere (Bora et al., 2016).

The SOC usually provides a framework for implementation and delivery of software agents, business process and logics over network based protocol. The SOC based model presents WS as loosely or tightly coupled software agents. It allows to present service as parent, broker or consumer of other service. Here, the consumer of service usually not aware of internal architecture of software agent being deployed. Their operational execution supports XML processing. The power of SOAP, UDDI and WSDL is being deployed for inter WS communication.

The WS can play the role of consumer, parent service (PWS) and child service (CWS). The WS can deliver services over network only through the web server. The web server processes the incoming Hyper Text Transfer Protocol (HTTP) request and delivers it to respective WS class files. Many researchers observed that (Saddik, 2006; Kalita et al., 2011; Bora et al., 2013; Bora et al., 2014), web servers have some limitations while handling massive request. Beyond a specific limit it fails for processing incoming request. However, the utilization of load balancing cluster based web server can enhance the processing capacity of web server up to a limit (Bryhni et al., 2000). The load balancing cluster based web server can be deployed through single or multiple machines. Different nodes can be created in the web server that can execute similar service for the end users. As such when a particular node fails in processing incoming HTTP request, the load balancer redirects it to other active nodes. The load balancer continuously validates whether the particular node is working with heavy load or idle in processing. As such

when a particular node fails or busy, the HTTP request is redirected to other nodes that are comparatively free in processing request. Thus, the overall capacity and performance is enhanced for service delivery. The deployment of this technique supports high availability, scalability and reliability against massive growth of end users. However, in the research community, identification stress limit beyond capacity, correlation and impact of different QoS metrics over WS execution and reliability beyond capacity of processing are gaining comparatively more importance.

OVERVIEW OF METRICS EVALUATION

Response Time

It is overall time required for processing the HTTP request and giving response to end users. The time unit is the parameter for estimation of service execution (Menasce et al., 2002).

Throughput

It highlights the data bytes process in server side. The measurement unit is bytes per seconds (Menasce et al., 2002).

CPU Utilization

It shows the utilization of processing resource during high stress of server execution. The measurement unit is usually the percentage of server execution (Singh et al., 2015).

Reliability

Reliability highlights the probability that the WS will deliver valid response to the consumer. The measurement is usually carried out through observation fault count within specific time interval. Many reliability models are there that can be used to analyse the fault tolerance of a service (Goel, 1985).

Related Work

Many researchers had carried out and presented their investigations over deployment of WS for computational implementation. However, there is a lack of study that specifically highlights the concern quality metrics of WS deployment during high stress of users.

Pacifici et al. had presented architecture for server overload and performance management of service during high stress of users (Pacifici et al., 2005).

Anzbock et al. had demonstrated a model for WS as middleware service deployment (Anzbock et al., 2005). They had evaluated a data sample to validate the WS execution.

Castro et al. had discussed a model for development of medical business process through different web technologies (Castro et al., 2006). However, the study lacks of experimental evaluation of the work.

Zhang et al. had presented a queuing model for performance aspects of load balancing web server (Zhang et al., 2008). Their experimental study highlighted the system metrics including response time and rejection rate of routing model for web server.

Medhi et al. had discussed a WS oriented Automatic Teller Machine (ATM) and shown some performance metrics using .NET framework (Medhi et al., 2014).

Kilintzis et al. had discussed the key aspects of deploying WS for data store operations. Their experimental results had validated the applicability of the work (Kilintzis et al., 2014).

Ismail et al. had discussed an approach for internal and external assessment of the service deployment (Ismail et al., 2015). An experimental study was carried out to establish the applicability of the work.

Medhi et al. had highlighted the reliability evaluation framework considering fault tolerance record of the WS using SOAP based Microsoft Windows Communication Foundation (WCF) service (Medhi et al., 2017).

Jiang et al. had demonstrated a study of reliability management for deployment of single service oriented WS (Jiang et al., 2017).

Abdelfattah et al. had discussed about reliable service architecture for ensuring service communication under different system constraint (Abdelfattah et al., 2017).

Zhou et al. had introduced a novel methodology for comparing security attributes and values for end users (Zhou et al., 2017).

Bora et al. had discussed failure and reliability aspects of SOC based system based on different deployment technique (Bora et al., 2018). The experimental study demonstrated the comparative key aspects of WS deployment using Tomcat and Microsoft IIS web server.

Yamada et al. had presented an analysis model to validate the smart grid based system through the aggregation of WS (Yamada et al., 2018).

Nato et al. had demonstrated a testing methodology for characterization of semantic WS (Nato et al., 2018).

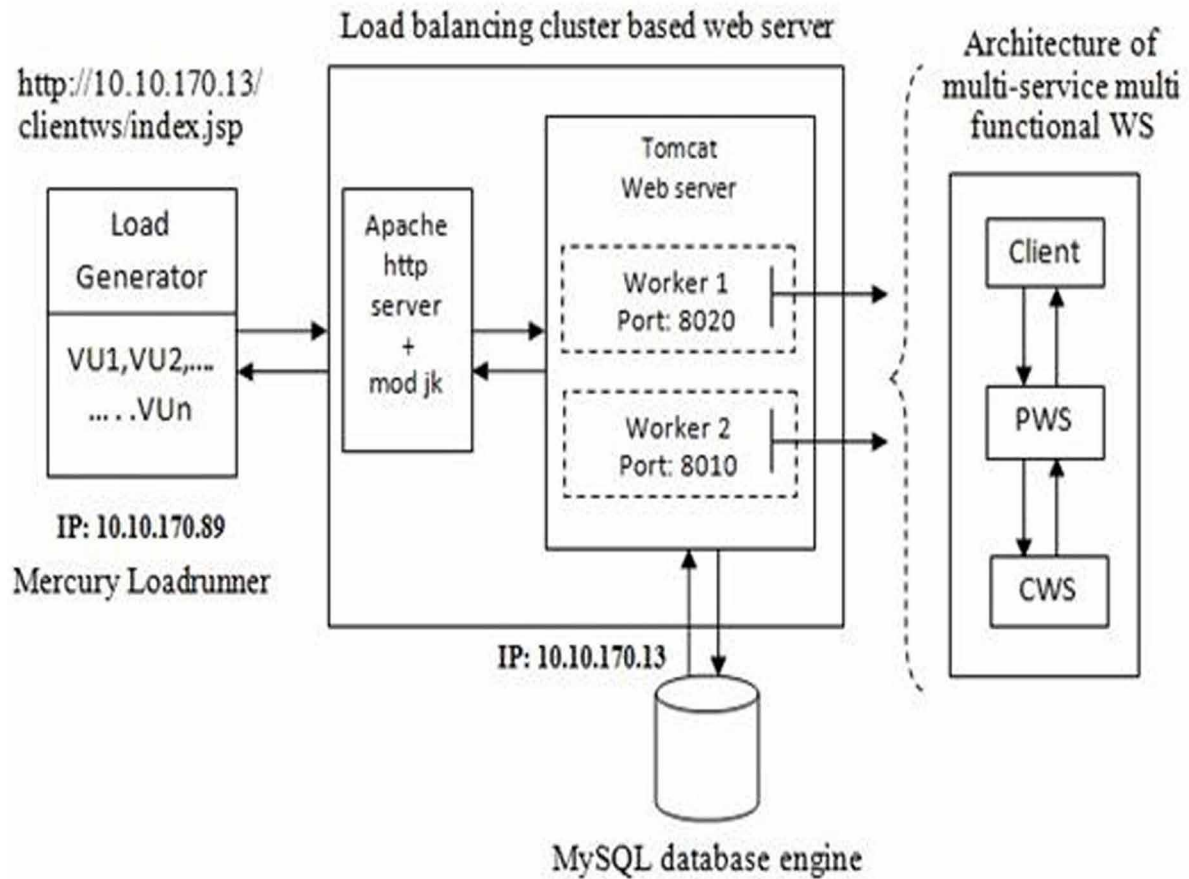
The proposed work is novel from precious work as it will discuss a QoS framework for evaluating the effectiveness of performance and reliability factors of SOC based system with multiple WS against massive load of request. The study will also highlight the correlation of system metrics such as response time, throughput, CPU utilization, load testing, fault analysis and reliability evaluation against different loads of user. The system metrics will be evaluated through statistical analysis to establish the effectiveness of WS in cluster based web server.

FOCUS OF THE ARTICLE

The focus of the work will be given on effectiveness of WS using load balancing cluster based web server. The key aspects of clustering different nodes of web server and deploying multi WS over there will be another focus for the work. The emphasize is given over the performance and reliability metrics of WS execution in cluster based web server as the end users visualizes them before utilization of service.

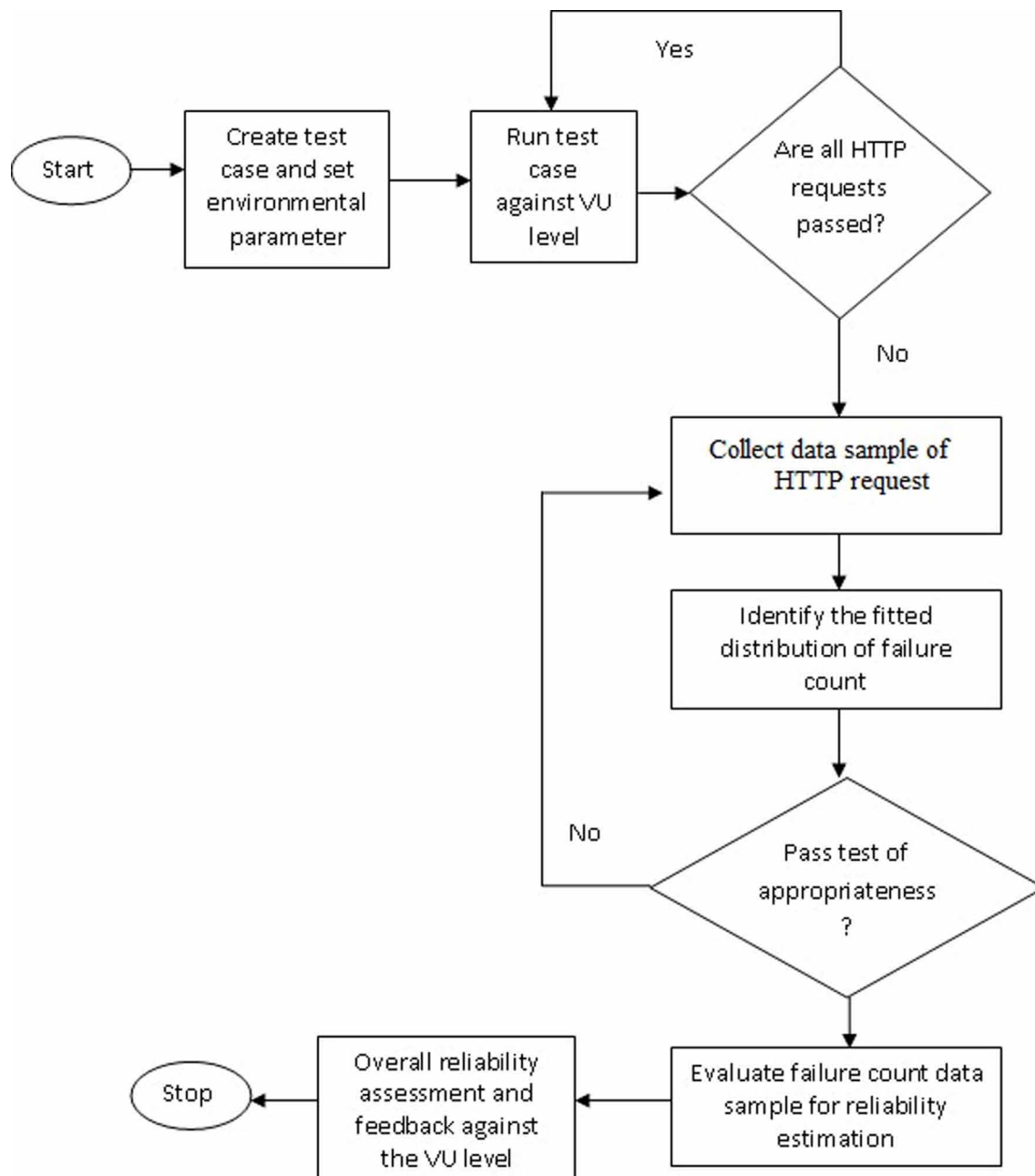
An architecture of multi WS communication is taken for the study that contains parent, broker and child WS and executes together to fetch record from database. The architecture of the service deployment is shown in Figure 1. In this study, the web server load is distributed in round robin scheduling fashion and the web server clustering is carried out through configuration of Apache Tomcat Application Server, mod_jk tomcat connector and Apache Http Web Server. The WS is developed and implemented using JAVA with Spring framework Model View Controller (MVC) version 2.5 and MySQL database server. A database mapping of 15000 records for disease medicine clinical remarks is taken for the study. The load in the system is given through the load testing tool Mercury LoadRunner version 8.1 that generates virtual users (VU) for accessing the system (Mercury LoadRunner, 2018). A flow chart for evaluating the effectiveness of the service is prepared and shown in Figure 2.

Figure 1. The architecture of service deployment



A test case is designed that contains instructions for the VU about how to access the system. The test case is discussed elsewhere at (Bora et al., 2014). All VU follows the test case. The Tomcat Application Server is clustered into different nodes running the same multi WS architecture. The load distribution of incoming request is handled by configuring Apache Http Web Server along with mod_jk tomcat connector. An extensive statistical analysis is performed over a data sample of recorded metrics considering response time, throughput and CPU utilization along with the failure density of the service. The hardware and software specification that will be used are: 64-bit Windows Server 2008 R2 Standard operating system (OS), Intel® Xeon® CPU E5620 @ 2.40 GHz processor, 8GB RAM, Apache Tomcat version 7 application server, Apache HTTP server version 2.2.4 web server, MySQL version 5.0 database server, 600 GB hard disk, Mod JK Tomcat connector, Java Development Kit of version 7, Java Runtime Environment of version 7, Spring MVC 2.5 framework, Metro WS stack, NetBeans Integrated Development Environment of version 7.0, Google Chrome web browser. The load will be generated using a machine with hardware and software specification as Windows XP OS, Intel® Pentium® Dual CPUE2200 @ 2.20 GHz processor, 1GB RAM, 150 GB hard disk and Mercury LoadRunner load testing tool of version 8.1.

Figure 2. Flowchart for evaluating the effectiveness of the service



SOLUTIONS AND RECOMMENDATIONS

Rigorous testing and deployment of statistical analysis are considered to be good candidature for evaluating the applicability and effectiveness of a system (Kalita et al., 2011). To study the effectiveness of the service, the SOC based system is evaluated against the load of 50, 100, 800, 1500, 1700, 1800, and 2000 VU level. The mercury load runner is deployed for generating different VU level. The necessary test case for accessing load balancing cluster based web server is created. All stress level of VU follows

Table 1. Recorded results against different loads of VU for two working node cluster based server (R_T : Response Time, s; T_{HPT} : Throughput, bytes/s; CPU_{UT} : CPU utilization, % ; TP: Transaction passed; TF: Transaction failed)

VU level	Average of Recorded parameters			TP	TF	Failure rate (%) = TF/(TF+TP)
	R_T	T_{HPT}	CPU_{UT}			
50	11.69	2368.013	3.88	287	0	0
100	11.86	4013.95	6.16	695	0	0
800	11.93	6394.106	19.246	19,566	0	0
1500	11.98	7324.544	21.094	62,968	0	0
1700	11.347	8592.85	22.161	80,040	0	0
1800	25.28	105729.1	39.13	82387	43836.1	0.35
2000	26.31	18251.2	43.6	91345	69251	0.43

the same script. Each execution of test case causes the event of establishing the client server communication through WSs. Each request causes the inter WS communication for processing BL. The test bed is set up in such a way that all VU will enter gradually the ramp up period of server and will execute together. During this ramp up period the measurement parameter such as response time, throughput and CPU utilization along with HTTP transaction passed and failed is being recorded through mercury load runner. However, after ramp up time, all VU exits the system simultaneously. The test bed is executed till the SOC based system throws HTTP transaction errors. Here, the failure records shows against the stress level of 1800 and 2000 VU. The failure record is observed to be increasing gradually beyond this limit. The effectiveness of the service against different stress level of VU is shown in Table 1. It is observed the recorded parameters are increasing with increase in stress level of VU.

Figure 3. Response time of the service against VU stress level of 100.

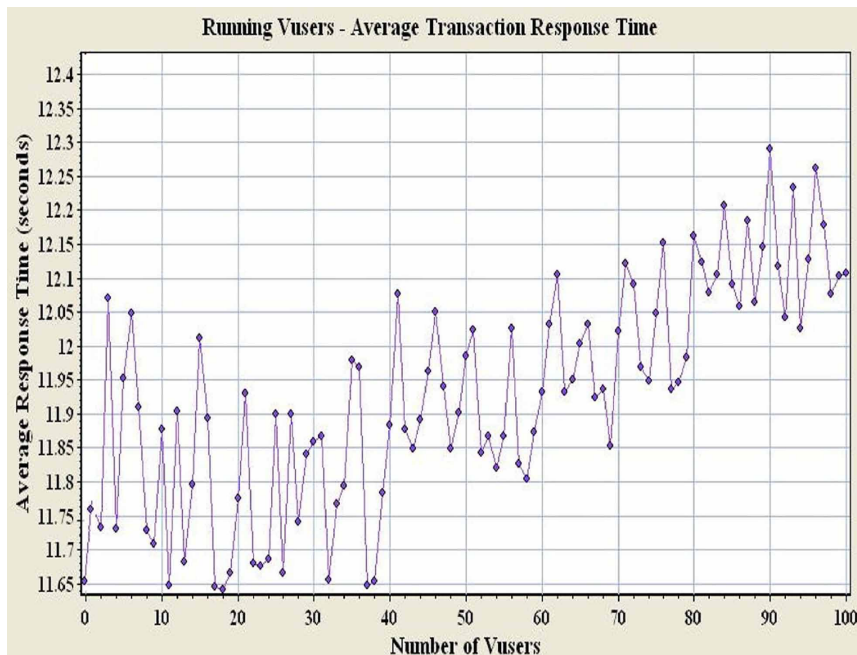
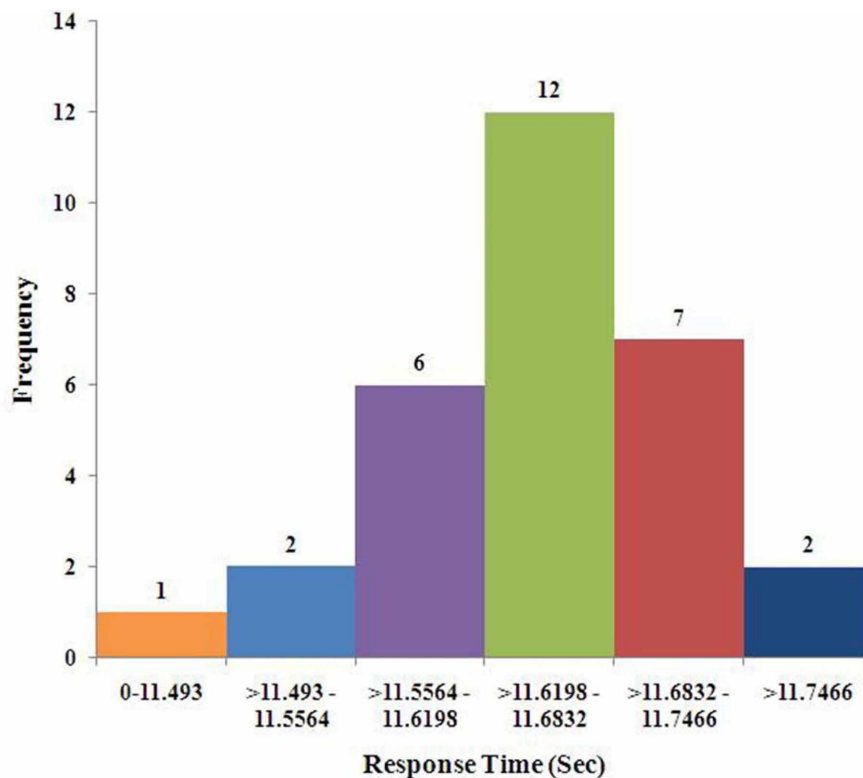


Figure 3 shows the recorded sample response time of the service against 100 VU. It is observed to be increased with the increase in stress level.

Distribution of Response Time

To study the distribution of response time of the service, a data sample of 30 repetitive tests against 50 VU stress level is taken of the study. The histogram is plotted against the data sample of recorded response time. The histogram is shown in Figure 4. It shows that the distribution follows normality with single peak value. However, based on different class sizes, different assumption may be interpreted. As such, the normal probability plot (NPP) is interpreted against the recorded response time. The NPP is a graphical interpretation technique that shows the normality of the recorded data points against theoretical quantile parameter (NIST, 2017). The NPP of the recorded response time is shown in Figure 5. The NPP of response time shows that the response time of the service follows a straight line. As such the recorded data sample can be interpreted to be linear and normally distributed. That means, the recorded data points are valid and follows the mean of the response time.

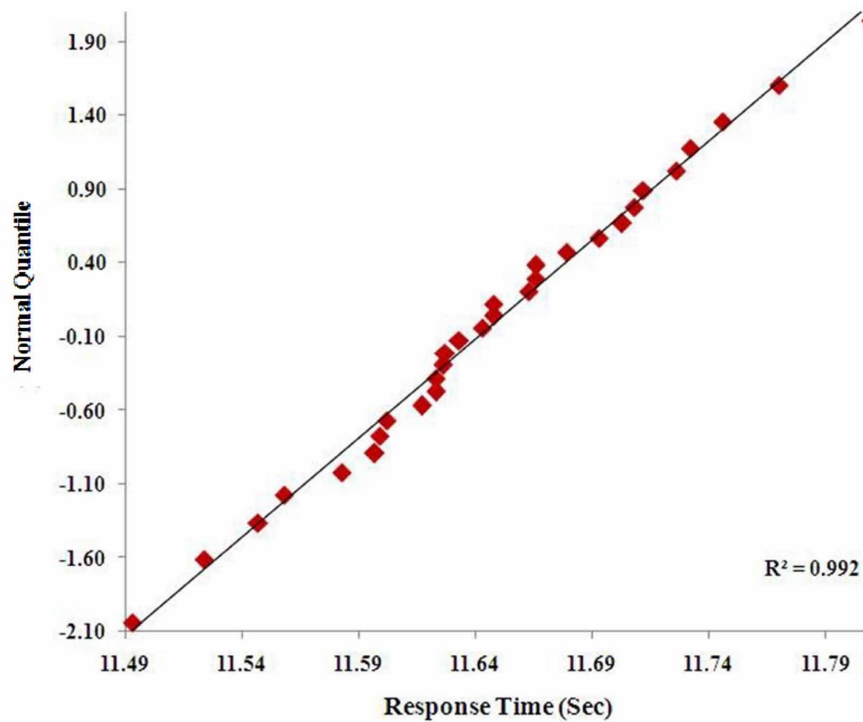
Figure 4. Histogram of response time



Correlation of Recorded Metrics

To study correlation of the recorded parameters, we deploy the regression analysis. A hypothesis (H_0) is considered that the response time, throughput and CPU utilization are correlated and the throughput

Figure 5. NPP of response time



and CPU utilization together have impact over the response time. At 95% confidence level, the study of multiple linear regression analysis is carried out and shown in Table 2.

Table 2. Multiple linear regression analysis of recorded data sample

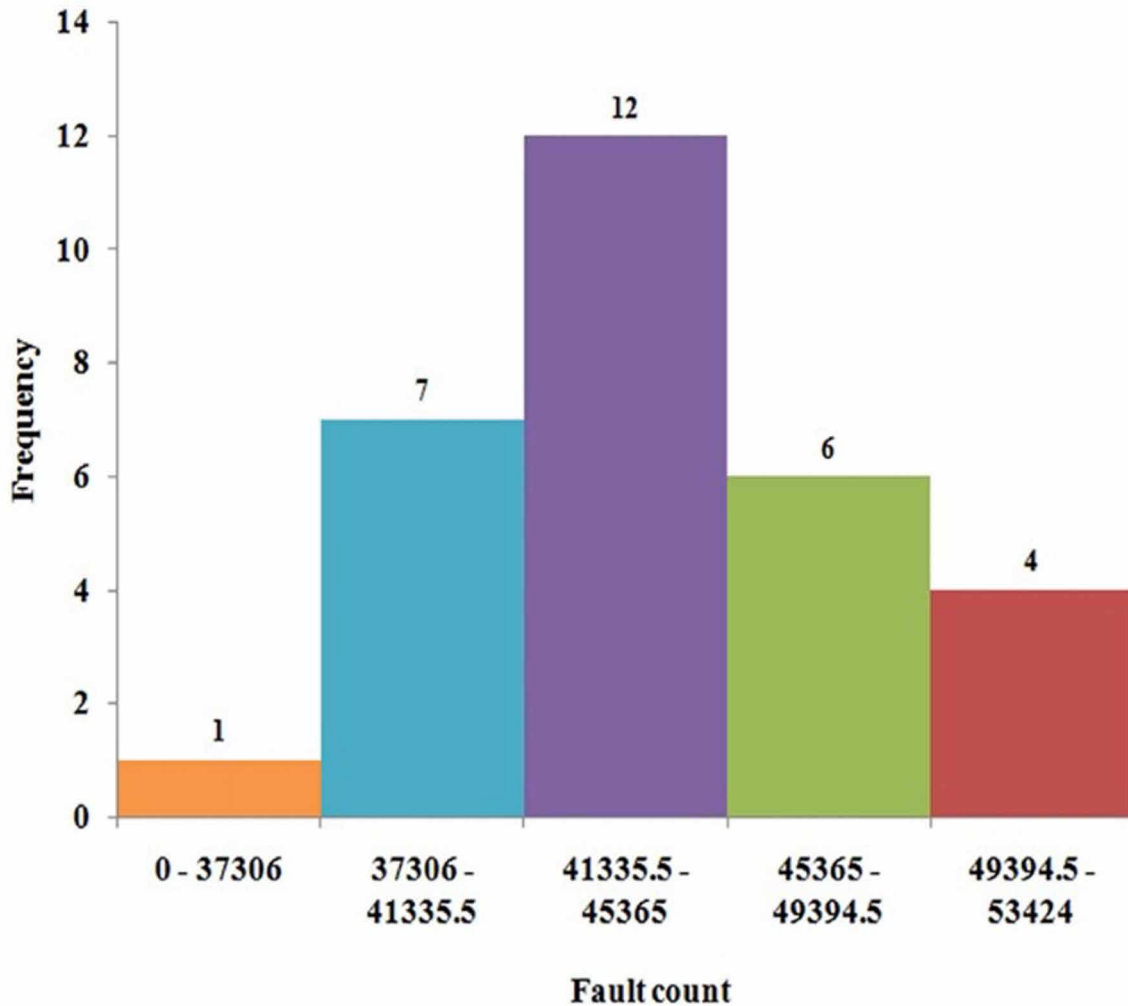
F ratio	F table critical value F(2, 27)	Is F ratio > F table critical value?	Adjusted R ²
5.4	3.354	Yes, Accept H ₀	23.3%

This validates the H₀ and concludes that recorded parameters are correlated and the response time is influenced by the recorded throughput and CPU utilization of SOC based system.

Evaluation of Reliability and Fault Count

The effectiveness of the reliability can be evaluated through the study of recorded HTTP fault count. The recorded experimental parameter as shown in Table 1 shows that the first HTTP failure occurs at the stress level of 1800 VU. As such, to study the mean fault count and reliability aspects of the service, we repeatedly tested the service against 1800 VU for 30 times. The histogram of the recorded fault count data sample is shown in Figure 6. It shows a mean HTTP fault count of 43836.1 and lies in the range of 41335.5 to 45365. The reliability of the service can be evaluated using equation (1) as defined by (Shooman, 2002, Rahmani et al., 2014, Medhi et al., 2017).

Figure 6. Histogram of recorded fault count



$$R = e^{-ft} \quad (1)$$

Here, the R metrics shows the reliability of the service. It is evaluated using failure rate 'f' as observed during 't' exposure period of the service. The R predicts the probability of valid execution of WS along with load balancing cluster based web server. That is, with valid value of R predicts that the WS will provide the searchable information during exposure period. Equation (1) evaluates a value within a range of 0 to 1. A value nearer to 1 predicts strong reliability. A service with poor reliability will have a value nearer to 0. In this case, 'f' value for the recorded 1800 VU data sample is observed to be 0.35%. As such, the R is evaluated and observed to be 0.7. As such strong reliability may be assumed with a probability that 70% of its exposure period will be served during invocations by 1800 users.

A comparative evaluation with other author's work is carried out to study the reliability and stability assessment of the proposed deployment. It is shown in Table 3.

It is observed from Table 3, that up to a stress level of 1700 VU, the failure rate 0. That means strong reliability can be observed during the stress. However, in case of (Saddik, 2006), the first failure rate is 60.28% for a stress of 100 VU. Similarly, in case of (Kalita et al., 2011), 50% of failure rate is observed

Table 3. A comparative assessment of the proposed deployment with other authors work

No. of VU	FR % of proposed study	FR % of earlier study done by other researchers		
		Saddik, 2006	Kalita et al., 2011	Medhi et al., 2014
50	0	Not Tested	0	0
100	0	60.28%	50%	0
500	Not Tested	87.9%	Not Tested	0
800	0	Not Tested	Not Tested	Not Tested
1000	Not Tested	92.06%	Not Tested	37%
1500	0	Not Tested	Not Tested	75%
1700	0	Not Tested	Not Tested	Not Tested
1800	35%	Not Tested	Not Tested	Not Tested

against 100 VU. Similarly, in case of (Medhi et al., 2014), 37% of failure rate is observed against 1000 VU. In the proposed methodology, the service is throwing first failure response against the stress of 1800 VU. As such it can be assumed that the proposed methodology of deployment gives better reliability and stability in comparison to other's deployment technique. It can also be concluded that efficiency, reliability and stability of multi WS paradigm with load balancing cluster based web server is better than other deployment technique. The expectation of service efficiency is also high during peak usage of the service. However, the system may generate failure record after maximal capacity of processing limit. As such SOAP request response processing error can generate and hamper the service delivery standards. The overall assessment concludes that segregation of role such as parent, broker and child service among different WS is better than merging the roles into a single WS.

FUTURE RESEARCH DIRECTIONS

As part of future work the proposed framework can be deployed for grid based system where WS will deliver services through independent machines. A proper fault tolerance methodology can be deployed for handling failure request and routing to the specific node for service delivery.

CONCLUSION

The work emphasizes the quality assessment framework for deployment of WS in load balancing cluster based web server. Deployment of different clustering nodes describes different aspects of effectiveness of the service against massive growth of users. The stability and reliability of the WS is observed up to 1700 load of users. The correlation of system metrics highlights that with increase in usage level, the system metric increases. A strong reliability can be assumed up to a specific load level of 1700 VU. Beyond that, moderate reliability can be assumed. Each stress generation through VU over the server causes SOAP communications among WSs. As such, with increase in stress level the response time, throughput and CPU utilization increases proportionally. The normality of data distribution is observed for the recorded data sample. This reveals that the population of the data sample follows the mean of recorded parameters. The effectiveness of the proposed work against high usage of the service is observed to be

better than other's deployment technique. This work precisely describes the quality evaluation framework along with the experimental results. As such, the software practitioners and researchers can get an in depth idea about the quality factors of WS against high load of users and the key metrics degrading the quality and reliability of the server. From the study and experimental results of the work the authors can conclude that the quality evaluation framework considering WS based system in load balancing cluster based web server will be viable, acceptable and applicable. The experimental results and evaluation of the work can help researchers and software developers to focus an idea about the quality of WS in load balancing cluster based web server.

ACKNOWLEDGMENT

The authors express their sincere thanks to All India Council of Technical Education (AICTE), Govt of India for financial support towards the work (Grant No. 8023/BOR/RID/RPS (NER)-84/2010-2011 31st March 2011).

REFERENCES

- Abdelfattah, A., Abdelkader, T., & El-Horbaty, E. (2017). RSAM: An enhanced architecture for achieving web services reliability in mobile cloud computing. *Computer and Information Sciences*, 30, 164-174. DOI: doi:10.1016/j.jksuci.2017.03.002
- Anzbock, R., & Dustdar, S. (2005). Modeling and implementing medical Web services. *Data & Knowledge Engineering*, 55(2), 203–236. doi:10.1016/j.datak.2005.03.009
- Bora, A., & Bezboruah, T. (2014). Testing and Evaluation of a Hierarchical SOAP based Medical Web Service. *International Journal of Database Theory and Application*, 7(4), 169–188. doi:10.14257/ijdt.2014.7.4.13
- Bora, A., & Bezboruah, T. (2016). *Some Aspects of QoS for High Performance of Service Oriented Computing in Load Balancing Cluster Based Web server. In Handbook of Research on Recent Developments in Intelligent Communication Application*. IGI- Global. doi:10.4018/978-1-5225-1785-6
- Bora, A., Medhi, S., & Bezboruah, T. (2018). *Investigations on Failure and Reliability Aspects of Service Oriented Computing Based On Different Deployment Technique. In Advanced Computational and Communication Paradigms. Lecture Notes in Electrical Engineering* (Vol. 475). Springer. doi:10.1007/978-981-10-8240-5_60
- Castro, V., Sanz, M., & Marcos, E. (2006). Business Process Development based on Web Services: a Web Information System for Medical Image Management and Processing. *IEEE International Conference on Web Services*, 807 – 814.
- El Saddik, A. (2006). Performance measurement of Web Service based application. *IEEE Transactions on Instrumentation and Measurement*, 55(5), 1599–1605. doi:10.1109/TIM.2006.880288
- Ismail, H., Issa, K., & Abdallah, K. (2015). Designing High Performance Web-Based Computing Services to Promote Telemedicine Database Management System. *IEEE Transactions on Services Computing*, 8(1), 47–64. doi:10.1109/TSC.2014.2300499

Jiang, P., Elag, M., Kumar, P., Peckham, S. D., Marini, L., & Rui, L. (2017). A service-oriented architecture for coupling web service models using the Basic Model Interface (BMI). *Environmental Modelling & Software*, 92, 107–118. doi:10.1016/j.envsoft.2017.01.021

Kalita, M., & Bezboruah, T. (2011). Investigations on performance testing and evaluation of PReWebD: A. NET technique for implementing web application. *IET Software*, 5(4), 357–365. doi:10.1049/iet-sen.2010.0139

Kilintzis, V., Beredimas, N., & Chouvarda, I. (2014). *Evaluation of the performance of open-source RDBMS and triplestores for storing medical data over a web service*. doi:10.1109/EMBC.2014.6944623

Matthew, M. C., Laskey, K., McCabe, F., Brown, P., & Metz, R. (2005). *Reference Model for Service Oriented Architectures*. Published on the internet. OASIS Working Draft.

Medhi, S., & Bezboruah, T. (2014). Investigations on implementation of e-ATM Web Services based on. NET technique. *International Journal of Information Retrieval Research*, 4(2), 42–51. doi:10.4018/ijirr.2014040103

Medhi, S., Bora, A., & Bezboruah, T. (2017). Investigations On Some Aspects of Reliability of Content Based Routing SOAP based Windows Communication Foundation Services'. *International Journal of Information Retrieval Research*, 7(1), 17–31. doi:10.4018/IJIRR.2017010102

Mercury LoadRunner. (n.d.). Available at: https://qageek.files.wordpress.com/2007/05/loadrunner_tutorial.pdf

Neto, J., Moreira, A., & Musicante, M. (2018). Semantic Web Services testing: A Systematic Mapping study. *Computer Science Review*, 28, 140–156. doi:10.1016/j.cosrev.2018.03.002

Pacifici, G., Spreitzer, M., Tantawi, A. N., & Youssef, A. (2005). Performance Management for Cluster-Based Web Services. *IEEE Journal on Selected Areas in Communications*, 23(12), 2333–2343. doi:10.1109/JSAC.2005.857208

SOAP. (n.d.) Retrieved online on 31.1.2016, Available at <http://www.w3.org/TR/soap12-part1/>

Yamada, T., Suzuki, K., & Ninagawa, C. (2018). Scalability Analysis of Aggregation Web Services for Smart Grid Fast Automated Demand Response. *2018 IEEE International Conference on Industrial Technology (ICIT)*. 10.1109/ICIT.2018.8352363

Zhang, Z., Fan, W. (2008). Stochastics and Statistics Web server load balancing: A queueing analysis. *European Journal of Operational Research*, 186, 681–693. DOI: doi:10.1016/j.ejor.2007.02.011

Zhou, B., Zhang, Q., Shi, Q., Yang, Q., Yang, P., & Yu, Y. (2017). Measuring webservice security in the era of Internet of Things. *Computers & Electrical Engineering*, 1–11. doi:10.1016/j.compeleceng.2017.06.020

ADDITIONAL READING

Bora, A., Bhuyan, M. K., & Bezboruah, T. (2013). Investigations on Hierarchical Web service based on Java Technique. In *Proceedings of the World Congress on Engineering (WCE)*.

Kalita, M., & Bezboruah, T. (2012). Investigations on implementation of web applications with different techniques. *IET Software*, 6(6), 474–478. doi:10.1049/iet-sen.2011.0136

Kalita, M., Khanikar, S., & Bezboruah, T. (2011). Investigation on performance testing and evaluation of PReWebN: A JAVA technique for implementing web application. *IET Software*, 5(5), 434–444. doi:10.1049/iet-sen.2011.0030

Medhi, S., Bora, A., & Bezboruah, T. (2016). Security Impact on e-ATM Windows Communication Foundation Services using Certificate based Authentication and Protection: An implementation of Message Level Security based on .NET Technique. *International Journal of Information Retrieval Research*, 6(3), 37–51. doi:10.4018/IJIRR.2016070103

Medhi, S., Bora, A., & Bezboruah, T. (2017). Investigations On Some Aspects of Reliability of Content Based Routing SOAP based Windows Communication Foundation Services. *International Journal of Information Retrieval Research*, 7(1), 17–31. doi:10.4018/IJIRR.2017010102

Rahmani, M., Azadmanesh, A., & Siy, H. (2014). Architectural reliability analysis of framework-intensive applications: A web service case study. *Journal of Systems and Software*, 94, 186–201. doi:10.1016/j.jss.2014.03.070

KEY TERMS AND DEFINITIONS

Load Balancing Cluster Based Web Server: It is a methodology where web servers can be clustered and loads among them can be balanced through load balancer server.

Load Testing: It is a testing candidature usually carried out to study the behaviour of a system against excessive load.

Service-Oriented Computing: It is framework for deployment of WS where we can configure and develop necessary WSDL, SOAP message structure, and BL method for both client and server side.

Test Case: It is a collection of instruction for the testing module. It highlights the way a tester will follow for testing the service.

WS Scalability: It is a term used to describe how scalable a system is. A WS is scalable if it can handle massive request efficiently.

WS Stability: It is a term used to describe how stable a system is. A WS is stable if it can execute without any error while processing incoming HTTP request.

Challenges and Opportunities of VLC Application in Intelligent Transportation Systems

Aleksandra Kostic-Ljubisavljevic

University of Belgrade, Serbia

Branka Mikavica

Faculty of Transport and Traffic Engineering, University of Belgrade, Serbia

INTRODUCTION

Visible Light Communication (VLC) is a developing wireless communication technology, which uses visible light (380 - 780 nm) for data transmission. It offers significantly greater bandwidth compared to the radio frequency (RF) communications. VLC exploits the low-cost and highly efficient Light-emitting Diodes (LEDs) used for both illumination purposes and wireless communications. LEDs are characterized by numerous advantages, including long lifespan, compact form factor, reduced usage of harmful materials in design and low heat generation even after long period of usage. Another important benefit of LEDs is the capability of switching to different light intensity levels (undetectable to human eyes). Furthermore, VLC is safe for human health and it does not affect the functionality of the sensitive electronic systems. It is also characterized by a high directivity and a predictable channel, which reduces the sources of interference and a high-security level is guaranteed due to the inherently reserved channels. These benefits make VLC a very promising technology for numerous applications. However, the high directivity requires the existence of the Line-of-Sight (LOS), which limits the use of VLC. Relative mobility in establishing communication links between vehicles or between vehicles and infrastructure can disrupt LOS. In addition, natural and artificial light also degrades performances of the VLC systems, especially the sunlight which adds noise and interference in the channel.

VLC enables wireless communications with high data rates wherever there is artificial lighting, indoor and outdoor. Thus, VLC has great potentials in vehicular communication applications. It can easily be incorporated into Intelligent Transportation Systems (ITS). The aim of ITS is to increase safety, the efficiency of the transportation system and to reduce the gasses emissions. Numerous studies show that VLC can satisfy the requirements imposed on vehicular networks. Various modes of vehicular communication can be established: Vehicle-to-Infrastructure (V2I), Vehicle-to-Vehicle (V2V) and Infrastructure-to-Vehicle (I2V). Considering the communications between infrastructure and vehicles, and vice versa, most of the existing systems are oriented towards the communication between LED-based traffic lights or street lighting system and vehicles. The high power of traffic lights allows relatively long communication distances. On the other hand, the short distance between the street light and vehicle, along with the high power implied, assures high data rates and increases the communication stability. Based on the LED lighting systems integrated into vehicles in headlamps, brake lights or signal lights, the VLC communication between vehicles is enabled. Besides communication, VLC enables positioning superior to the Global Positioning System (GPS) with lower costs than existing systems. In this chapter, the

DOI: 10.4018/978-1-7998-3479-3.ch072

comprehensive analysis of VLC technology and applications in ITS are provided. Main characteristics of VLC in ITS in terms of architecture, modulation and standardization are analyzed. Benefits and possible challenges of VLC application in ITS scenarios are also addressed.

BACKGROUND

Transportation systems play an important role in modern life. The advancements in communication technologies and their application in transportation systems have led to the introduction of the concept of Intelligent Transportation Systems (ITS). A subset of the ITS, in which participants communicate and share information is called Cooperative Intelligent Transportation Systems (C-ITS). C-ITS aims at improving safety, sustainability and efficiency. C-ITS consists of the following four entities: personal, vehicle, central and roadside (ETSI EN 302 655 v.1.1.1, 2010). Personal entity enables access to ITS applications using personal devices, such as Smartphone with location-based applications. Vehicle entity consists of the vehicle on-board equipment (on-board unit, OBU) which hosts ITS applications. These vehicular ITS applications gather information about the vehicle and its environment and provide partial or full control over the vehicle in critical situations. Central entity comprises the equipment used to maintain, monitor and provide functionality to ITS applications. Finally, roadside entity, also referred to as Roadside unit (RSU), provide the equipment placed along the roadside which hosts ITS applications. The roadside ITS applications communicate with the vehicular ITS sub-systems together information about road environment, traffic conditions and control roadside equipment. A Multi-tier ITS architecture is proposed by Lin, Wang and Ma (2017). This structure provides guarantees in terms of interconnection, operation and service. It consists of the physical layer, the communication layer, the operation layer and service layer. Physical layer comprises vehicles, infrastructure and end users. Communication layer enables data transmission between ITS elements. Communication options provided by the communication layer include Field-Vehicle Communications (communications between vehicles and infrastructure), Fixed Point-Fixed Point Communications (communications between stationary entities), Vehicle-Vehicle Communications (short-range communications between vehicles) and Wide Area Wireless (Mobile) Communications (communication between vehicles and end user mobile devices). In general, ITS applications rely on Vehicle-to-Infrastructure (V2I), Vehicle-to-Vehicle (V2V) and Infrastructure-to-Vehicle (I2V) communications. There are three main classes of ITS applications: Road/Traffic Safety, Traffic Efficiency and Value-added applications (ETSI EN 302 655 v.1.1.1, 2010). Traffic safety applications aim at minimizing the risk of undesirable situations and reducing the damage of unavoidable accidents. The improvement of traffic conditions (in terms of travel time and traffic congestion) is a primary goal of traffic efficiency applications. Traffic information is usually gathered by roadside infrastructure and transmitted to the user. Value-added applications include providing Internet access, travel information and planning. The goals of these applications are comfort and convenience for the end users.

ITS involves cooperative wireless communication technologies to provide data exchange between vehicles or between vehicles and infrastructure. Current ITS research activities, products and standardization are mainly based on radio frequency (RF) communication technologies. Due to numerous advantages, VLC is considered as a promising candidate technology for vehicular communications (Khan, 2017). The primary advantage of VLC is the possibility of deployment in vehicular environments using existing infrastructure, which makes VLC a low complexity and low-cost technology. Highly directional Line-of-Sight (LOS) enables positioning with high precision. The positioning error is up to tens of centimetres, which is more accurate than the RF-based positioning technology (Armstrong, Sekercioglu

& Neild, 2013). In addition, VLC systems are characterized by high scalability (You, Chang, Lin & Tsai, 2013). In the periods of heavy traffic densities, RF communication systems experience undesirable packet collisions, longer delays and poor packet reception rate. On the other hand, VLC systems allow only communication between vehicles that reach LOS. Thus, signal congestion is significantly reduced. Furthermore, VLC enables high data rates, due to vastly available bandwidth. The potential of VLC as a complement or an alternative to RF technologies in vehicular communications is analyzed by Cailean, Cagneau, Chassagne, Popa and Dimian (2014). A comprehensive comparison of RF and VLC in vehicular communications in terms of conditions in communication channel: physical waves, radiation pattern and path loss modeling, noise and interference and performance modeling is provided by Cheng, Viriyasitavat, Boban and Tsai (2017). A VLC-based I2V communication system is analyzed by Kumar, Lourenco, Teraa, Alves and Aguiar, (2012), where a LED-based traffic light is used as an RSU for traffic information broadcast. VLC based V2V communication system with an LED-based transmitter and a camera receiver is proposed by Takai, Harada, Andoh, Yasutomi, Kagawa and Kawahito (2014). Another system based on LED transmitter and camera receiver convenient for V2V and V2I applications is proposed by Takai, Ito, Yasutomi, Kagawa, Andoh and Kawahito (2013). This system achieves the data rate of 10 Mbps per pixel or more for concurrent sending of video data and vehicle internal data, such as vehicle speed, braking status, etc. In addition, this system enables real-time detection (Takai, Ito, Yasutomi, Kagawa, Andoh & Kawahito, 2013). Luo, Ghassemlooy, Le Minh, Bentley, Burton and Tang (2014) analyzed VLC V2V communication models based on LOS and Non-Line-of-Sight link configurations. The model determines the relation between the communication range and the system Bit Error Rate (BER). It appears that placing a photodetector at a height of 0.2 - 0.4 m above the road surface in the vehicle, the communication range can be lengthened up to 20 m at a data rate of 20 Mbps. Cooperation between VLC based I2V and V2V communications is observed by Cailean, Cagneau, Chassagne, Topsu, Alayali and Dimian (2013). Node to node communication is used to limit the interferences caused by multiple transmitters in the communication range. That system achieves BER of 10^{-7} for the both I2V and V2V ITS segments. In addition, it is possible to maintain low BER while increasing the distance between the transmitter and the receiver.

VLC outdoor applications need to solve several challenges (Vappangi & Mani, 2019). In general, the dynamic nature of the vehicular environment and the weather conditions cause that the VLC channel is very unpredictable. The major problem for outdoor VLC application is triggered by the numerous sources of parasitic light. Besides the artificial light sources, the sunlight represents the most powerful source of the noise. VLC channel conditions can be further deteriorated by snow, dust, fog or raindrops. Any obstacle that can obstruct the communication path affects the power of the received signal. Another important challenge for VLC implementation in vehicular applications is to enable data transmission at large distances. More efficient photosensitive elements need to be integrated to enable low-power signals detection, and hence, to enable longer communication ranges. Therefore, increasing the robustness to noise is an important issue to be solved for vehicular VLC applications.

KEY CHARACTERISTICS OF VLC IN ITS SCENARIOS

In this section, the architecture, modulation techniques and standardization issues of VLC in the segments of ITS are analyzed.

A VLC system comprises the VLC transmitter and the VLC receiver. The VLC transmitter is an optoelectronic transducer that modulates the light from a LED luminaire. The LED luminaire consists

of a LED lamp, ballast, housing and other components. The LED lamp consists of one or more LEDs and a driver circuit which controls the brightness. When a LED luminary is used for communications, the driver circuit is modified to enable modulation. Since the VLC systems are used for both illumination and communication, the major requirement in the VLC systems is that illumination should not be affected due to communication. The VLC receivers can be classified into two types: photodetector (also referred to as photodiode or non-imaging receiver) and imaging sensor (also referred to as camera sensor). The photodetector is a semiconductor device that samples the received light and converts it into electric current. An imaging sensor requires a lot of photodetectors for high-resolution photography. Very-low-data rates can be achieved using imaging sensors. However, the data rate can be improved using the “rolling shutter” effect (Pathak, Feng, Hu & Mohapatra, 2015).

There are three possible VLC link realizations: directed LOS, non-directed LOS and diffused non-LOS. The directed LOS requires the precise alignment between the VLC transmitter and the VLC receiver. Therefore, it enables the highest throughput and the longest distance. In the case of the non-directed LOS, the requirements in terms of the alignment are less strict. However, only shorter distances can be achieved and medium to high throughput is provided. The diffuse non-LOS does not impose any alignment requirements. As a result, the lowest throughput is provided and it is convenient only for indoor use.

The light signal is characterized by the non-encoding properties of the data in phase and amplitude. Therefore, the data in VLC systems need to be encoded using variations in the intensity of the light emitting wave. It is important to emphasize that any modulation technique in VLC should not only achieve higher data rate but also satisfy the requirements in terms of dimming and flickering of perceived light to humans (Pathak, Feng, Hu & Mohapatra, 2015). Different levels of illuminance are needed for different environments. Thus, an illuminance in the range of 30 - 100 *Lux* is convenient for public places. Indoor applications require higher illuminance. Nowadays, LEDs support dimming at an arbitrary level depending on the requirements and hence, it is possible to achieve significant energy savings. The fluctuations in the brightness of the light, i.e. flickering effect can be very harmful to human. Therefore, the light intensity should change at a rate faster than the human eye can perceive. According to IEEE 802.15.7 (2011), flickering should be faster than 200 Hz.

The following modulation techniques are commonly used: On-Off Keying (OOK), Pulse modulation techniques and Color Shift Keying (CSK).

OOK modulation technique represents the transmission of bits “1” and “0” by turning the LED on and off, respectively. During the OFF state, the light intensity is reduced. Major advantages of OOK are simplicity and ease of implementation. The white LEDs are most commonly used for OOK modulation in VLC. It is possible to achieve 10 Mbps using Non-Return-To-Zero (NRZ) OOK with the white LED (Park, Jung, Shin, Hyun, Lee & Oh, 2007). A blue filter can be used to enhance the data rate up to 40 Mbps. In addition, the blue-filtering with analogue equalization at the receiver enable data rate of 125 Mbps (Vucic, Kottke, Nerreter, Habel, Buttner, Langer & Walewski, 2009). The performance can be further improved by using an avalanche photodiode at the receiver. The combination of the avalanche photodiode and NRZ-OOK enables data rates up to 230 Mbps (Vucic, Kottke, Nerreter, Habel, Buttner, Langer & Walewski, 2010). The data rates up to 477 Mbps can be achieved using RGB white LED and NRZ-OOK modulation, where only the red LED is modulated for data transmission while the other two are provided with the constant current for illumination.

Various modulation techniques based on pulse width and position are used to enable higher data rates, especially when supporting different dimming levels. In Pulse Width Modulation (PWM) the widths of the pulses are different for different dimming levels. The advantage of PWM is that enables dimming without changing the intensity level of pulses. Hence, the colour shift in the LED is avoided. The major

limitation of PWM is the low data rate of up to 4.8 Kbps (Sugiyama, Haruyama & Nakagawa, 2007). In Pulse Position Modulation (PPM), the symbol duration is divided into slots of equal duration and a pulse is transmitted in one of the slots. The position of the pulse determines the transmitted symbol. Due to limitations of low spectral efficiency and low data rate, variations of PPM technique are proposed. Overlapping PPM (OPPM) allows more than one pulse to be transmitted during the symbol duration, while the different pulse symbols can overlap (Shiu & Kahn, 1999). Compared to PPM and OOK, OPPM achieves higher spectral efficiency and a wide range of dimming levels, along with the high data rate (Bai, Xu & Fan, 2010). Multipulse PPM (MPPM) also allows multiple pulses to be transmitted during the symbol duration, but the pulses within a symbol duration do not have to be continuous. Compared to OPPM, MPPM improves spectral efficiency (Sugiyama & Nosu, 1989). Spectral efficiency can be further improved using a combination of OPPM and MPPM, referred to as Overlapping MPPM (OMPPM). OMPPM enables that more than one pulse positions are allowed for each optical pulse. Moreover, Expurgated PPM (EPPM) is a modulation technique where symbols in the MPPM are expurgated to maximize the inter-symbol distance. EPPM provides dimming support as it can achieve an arbitrary level of illumination by changing the number of pulses per symbol and the length of the symbol. In addition, EPPM mitigates the flickering. Multi-level EPPM (MEPPM) supports multiple amplitude levels in order to improve the constellation size and spectral efficiency. It also supports dimming and flickering-free communication. A hybrid modulation technique comprising PPM and PWM is referred to as Variable PPM (VPPM). VPPM allows that bits are encoded by the different position of the pulse, while the width of the pulse can also be modified.

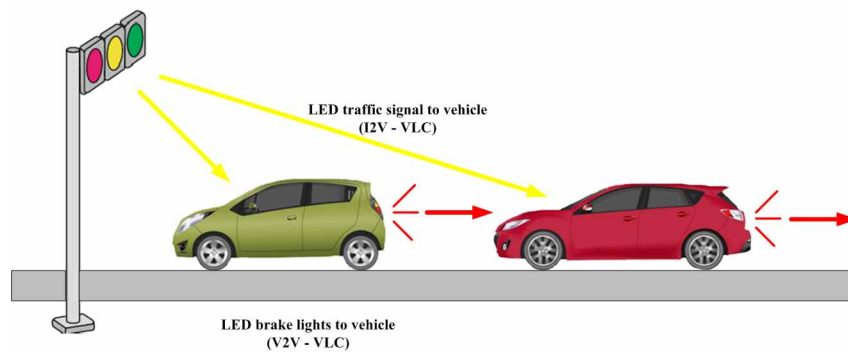
The white light can be generated using three separate LEDs, Red, Green and Blue (RGB), often referred to as TriLED (TLED). CSK modulation technique modulates the signal using the intensity of the three colours in the TLED source. Compared to other modulation techniques, CSK improves the data rate and it is based on the colour space chromaticity diagram (Pathak, Feng, Hu & Mohapatra, 2015).

Standardization initiatives are of great importance for further VLC integration in ITS scenarios. Two organizations are in charge of VLC standardization: the Visible Light Communication Consortium (VLCC) in Japan and the IEEE. The Japan Electronics and Information Technology Industries Association (JEITA) CP-1221, JEITA CP-1222 and JEITA CP-1223 are published by the VLCC. These standards mainly consider position detection applications. However, they are not commercially exploited. The first official VLC standard is provided by the IEEE 802.15.7 (2011). The aim of the standard IEEE 802.15.7 is to provide the access to several hundreds of THz bands, to prevent electromagnetic interference and to provide physical layer specifications, such as optical mapping, flickering and dimming mitigation etc. This standard should also support applications related to VLC Personal Area Networks (VPAN). It comprises both the physical layer (PHY) and the medium access control (MAC) layer for short-range optical wireless communications using visible light in optically transparent media. There are three types of the physical layer: PHY I, PHY II and PHY III. These layers can employ different modulation techniques and different data rates can be achieved. Thus, PHY I, PHY II and PHY III enable data rates in the range 11.67 - 266.6 kbps, 1.25 - 96 Mbps and 12 - 96 Mbps, respectively. For most ITS applications, PHY I type is the most convenient, since it is designed mainly for outdoor applications. PHY I typically uses OOK and VPPM modulation techniques. Although it provides the lowest data rates, it is possible to eliminate the additional path loss due to longer distance and potential interference from daylight and fluorescent lighting (Falcitelli & Pagano, 2016).

POSSIBILITIES OF VLC APPLICATION IN ITS

Integration of VLC systems in ITS architecture can play a significant role in the improvement of various aspects of transport and traffic engineering (Kostic-Ljubisavljevic, Radonjic Djogatovic & Mikavica, 2017). The transportation lighting infrastructure, such as street lighting, traffic lights, vehicular lamps, etc., are switching to LED lighting. This is a huge potential for low-cost VLC integration into the ITS environment. RSUs such as LED-based traffic lights can be used for information broadcast, as shown in Figure 1.

Figure 1.



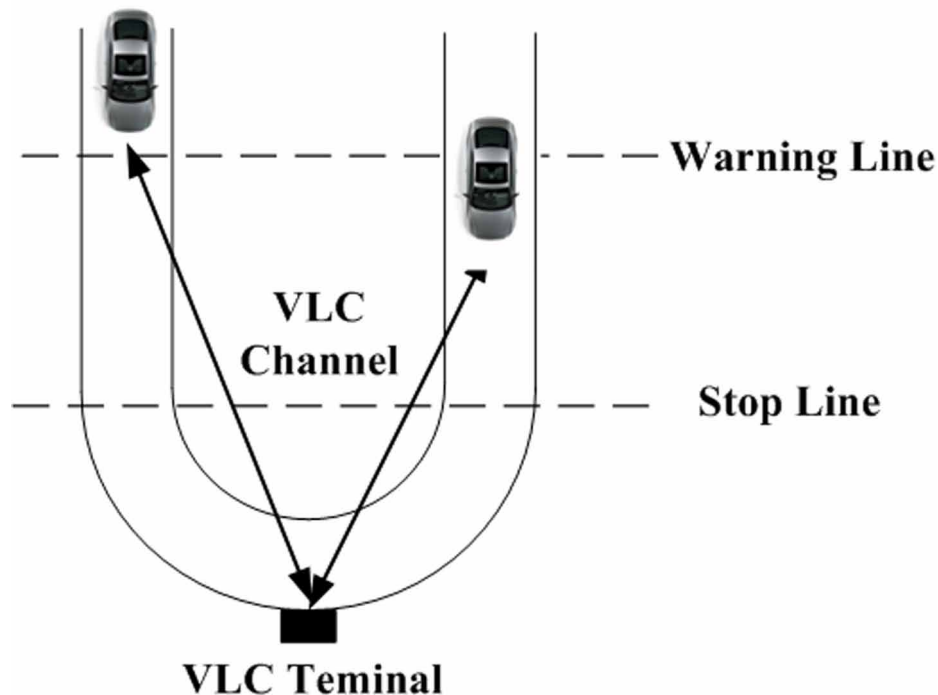
This information is then detected by receivers installed in vehicles, thus enabling I2V communications. Therefore, traffic safety information can be broadcasted without additional energy consumption. Furthermore, this information can enhance traffic flows and reduce accidental situations, as well. Due to straight light propagation, it is possible to achieve high-directional communications. Thus, different traffic information can be transmitted for different lanes of a road. Vehicles can also communicate with adjacent vehicles using head, tail or brake lights, thus enabling V2V communications. Potential applications of VLC-based V2I and V2V communication systems include active road safety, traffic efficiency, local services and Internet-based services. More VLC applications in ITS are explained below.

Traffic safety can be significantly improved by VLC based road surface detection. Three mechanisms occur when the light beams the surface: reflection, transmission and absorption. The diffused reflection is the reflection of light from a rough surface when an incident ray is being reflected in many angles. Some rays are reflected back around the angle of the incident ray. When the road surface is wet, two scenarios are possible. At first, water fills holes in the surface and enables a specular diffusion when an incident ray is being reflected at one angle. Thus, the diffused reflection is minimized. The other situation occurs when the water level is deep enough to provide a total internal reflection within the water puddle. Therefore, the diffused reflection depends on the water depth on the surface. Every change in the water level on the surface can be measured using a LED fog lamp to beam the surface and then receive the reflection from a photodetector placed within the headlight compartment (Cahyadil, Kim, Chung & Ghassemlooy, 2015).

Combining VLC-based V2V communications and visible-light positioning can improve vehicle safety systems. VLC lights can concurrently provide illumination and signaling, reliable communications and accurate positioning (Yu, Shih, Tsai, Wisitpongphan & Roberts, 2013). Such a system is characterized by low complexity and it is scalable in high vehicle density. Each vehicle in this system can create a map

that contains the relative positions, speed, brake status, and heading of the surrounding vehicles that reach LOS. Using the map with the estimated positions of surrounding vehicles, the driver can adjust the vehicle's speed or heading to avoid accident situations. Information about stopped and overtaking vehicles are also provided. Furthermore, the trajectories of the surrounding vehicles in the next few seconds can be estimated. Another system for traffic safety improvement provides a map with positions and speed of all surrounding vehicles, including those in nearby lanes (Yu, Shih, Tsai, Wisitpongphan & Roberts, 2013). Thus, when the driver indicates the willingness to change a lane, the system analyzes the map and signalizes when the lane change can be executed safely. This application requires very accurate positioning to identify the lane of each surrounding vehicle, which VLC can assure. Collision risk is also high at road curve lanes. In situations where the curve width only allows one vehicle to pass through each time, a VLC terminal equipped with LED indicator can be located outside the road curve lane to communicate with vehicles on both sides (Qiao, Xia, Hong & Shen, 2017). Whenever one vehicle is going into the curve lane, an alarm signal should be triggered and released to indicate another vehicle to reduce the speed or to stop. In such a system, as shown in Figure 2, a stop line and a warning line should be set. Thus, the vehicle first reaches the warning line and send the VLC signal to the VLC infrastructure terminal. Then, the VLC terminal broadcasts warning signals in the area that covers the whole curve lane. When the vehicle approaches the stop line, if there comes another vehicle at the same time it receives the warning signal to decrease the speed or to stop.

Figure 2.



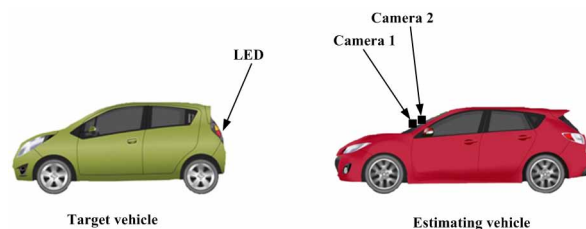
Positioning using VLC becomes an active research topic, especially for traffic safety improvement. Global Positioning System (GPS) lacks in provisioning accurate location information in high-density traffic conditions or in the city zones with high surrounding buildings. Outdoor VLC positioning systems

provide vehicles and pedestrians more accurate location service. In an outdoor environment, the main light source can be traffic lights, vehicle headlights/taillights, or street light posts. Outdoor applications are exposed to some challenges. At first, the lights can be poorly distributed or moving with the presence of ambient noise. Furthermore, real-time positioning of fast-moving vehicles is highly challenging. However, VLC based outdoor positioning has many advantages and main applications are described as follows.

Besides their main function as a signaling device, LED-based traffic lights can be used for continuous provisioning of positioning service for vehicles in its vicinity. The outdoor positioning can be performed using one or two traffic lights and two photodiodes (Bai, Chen, Xu & Fan, 2011). The signal, containing the light position information, is emitted from the traffic light. Two photodiodes are mounted on the front lights of the vehicle. Their centre is the required position of the vehicle. In the single-traffic light scenario, two photodiodes perform measurements of Time Difference of Arrival (TDOA) of the traffic signal when the vehicle is moving to another location. These measurements create two sets of the hyperbola. The observed vehicle is located within the area formed by the crossed points of these hyperbolas. The scenario with two traffic lights involves a vehicle traffic light and a pedestrian traffic light. Each light is modulated in different frequencies. Two photodiodes measure the TDOA simultaneously. Based on these measurements, it is possible to obtain the relative position of the traffic lights to the vehicle and the absolute position of the vehicle.

VLC system based on an image sensor can be used for V2V positioning, as shown in Figure 3 (Tram & Yoo, 2018). Cameras are embedded in the estimating vehicle to capture the targeted vehicle's LED. The distance between the observed vehicles is determined by capturing images. The LED is easy to detect on the image, due to the high illumination and compact shape. Using such a system, the positioning information can be transmitted between vehicles.

Figure 3.



VLC based positioning can also be useful in tunnels, where GPS has poor performances. In such a system, LED lighting infrastructure in a tunnel and LED tail lamp in a vehicle have the absolute position coordinates. Location data are transmitted using VLC based V2V and V2I communication systems. It appears that three LED light sources on the tunnel infrastructure and vehicle tail lamps based on the image sensor can provide accurate positioning with approximately 1m error (Kim & Jung, 2016).

SOLUTIONS AND RECOMMENDATIONS

VLC applications in ITS are facing some technical issues and challenges that can limit the communication range. Relative mobility between vehicles or between infrastructure and vehicles can hinder the LOS. This issue can be tackled by optimizing lighting positioning. In addition, natural and artificial lights

add noise and interference, which can be minimized by using optical filters and optimized electronics. Reliable communication is possible if the distance between a VLC transmitter and a VLC receiver is in the range of 40 - 50 m (Liu, Sadeghi & Knightly, 2011).

VLC systems in vehicular communication need to fully support the mobility of the vehicles. Studies related to VLC connection between moving vehicles mostly analyze the scenario where the transmitter and the receiver are relatively aligned. However, in real traffic situations, the alignment between the transmitter and the receiver is not always satisfied. A possible solution for this challenge is to use more photodetectors oriented for different reception angles. Another possible solution is to implement more than one sensor on each side of the vehicle. Thus, at least one sensor can be properly aligned with the transmitter and the communication link can be established.

VLC can be a very promising solution for vehicle platooning, due to relatively short communication link requirements and the type of the transmitted data. However, the presence of ambient noise significantly affects platooning. Exchanged information about the relative position and front and rear facing directions of each vehicle of the platoon can be processed in each vehicle controller (Abualhoul, Marouf, Shagdar & Nashashibi, 2014). In addition, it is possible to compensate the trajectories influence on the optical incidence and irradiance angles efficiently. Furthermore, the total optical path loss between any two members of the platoon is reduced.

In vehicular VLC systems, it is very important to achieve high data rates with respect to lighting quality and its correlation with dimming utility (Ucar, Turan, Ergen, Ozkasap & Ergen, 2016). This is crucial especially in night conditions when a high beam of headlights can create glare and cause temporary blindness of a driver. In order to avoid these unsafe situations, light sources should be arbitrarily dimmed. However, efficient dimming control is achievable at the expense of decreasing data rates. In addition, external weather conditions such as snow, rain and fog affect dimming. Therefore, efficient dimming functionality requires a compromise between illumination and communication. A comprehensive analysis of dimming techniques can contribute to traffic safety improvement and provides full control over the lighting output.

Compared to RF communication, multipath propagation in V2V VLC systems is negligible. The signal power of the multipath components is significantly reduced and these multipath components usually can be ignored. Furthermore, visual obstacles can block the propagation of multipath components. Hence, the characterization of V2V VLC channels comprises the path loss and the path loss variation over time (Cui, Wang & Tsai, 2014). The path loss depends on the distance between the transmitter and the receiver, the irradiance angle and the incidence angle. A major source of time variation of the received signal power is the mobility of the vehicle. The location of the tail lights of neighboring vehicles with respect to the receiving vehicle can be estimated using the video data from camera sensors mounted on vehicles. Knowing the estimated location parameters of a single image, the path loss of the link between the observed vehicles at that time instance can be estimated. This procedure can be performed in multiple series, thus determining the time function of the path loss. These data can be further used to characterize the channel fading due to vehicle mobility.

VLC has numerous advantages in vehicular applications. However, the establishment of LOS is a major disadvantage that limits performances. Since VLC and RF are complementary technologies, their integration can enhance mobility and increase the reliability of the connection. For instance, integration of VLC and Dedicated Short Range Communication (DSRC) technology can significantly improve performances in safety applications. DSRC is a mature technology that enables long-distance communication. However, VLC is not able to provide comparable communication ranges but has a huge potential in high traffic densities. Such heterogeneous networks in vehicular communications are considered as a promising solution for future vehicular networks.

FUTURE RESEARCH DIRECTIONS

Visible light communication has the potential to provide low-cost high-data rate communication with improved energy efficiency. However, there are some challenges and open issues that need to be considered in the near future in order to enable the widespread use of VLC in ITS. The main assumption in the techniques of achieving high data rates in VLC links is the presence of LOS, where the transmitter and the receiver have aligned the field of views. In practice, receiver movement and orientation changes are very often. Therefore, it is essential to develop methods that can provide high data rates even in the presence of the field of view misalignments. This task is very challenging and represents an important direction for future research. In addition, it is necessary to develop receivers that can support robustness to device movements and field of view misalignments. Such receivers should also run with low energy consumption. Besides the field of view alignment, another important issue is tackling the shadowing effects. Hence, new techniques that provide real-time adjustment depending on the changes in received signal power, due to the field of view misalignments and shadowing, need to be developed.

Flickering, i.e. the fluctuations in the brightness of the light, can have harmful effects on human health. If a data rate is slow, or idle period in the system is significantly longer compared to the period of data transmission, the flickering effect is more noticeable. Therefore, the modulation technique needs to be chosen appropriately, in order to avoid flickering. Designing a modulation technique for VLC systems that reduces flickering is a very challenging issue and requires a lot of research effort. The suitability of indoor modulation techniques for outdoor scenarios can also be a subject for future research directions. Another important open issue is dimming. Different levels of illuminance are required in different indoor and outdoor scenarios. The algorithms for dimming in VLC systems should be designed so that the requirements for illuminance in a given scenario are satisfied.

CONCLUSION

Visible light communication is an optical wireless communication technology which uses visible light as a communication medium. The VLC system provides highly direction communications in LOS scenarios in short range with high sensitivity to weather conditions and ambient noise. Main characteristics of VLC systems are the following: there is no health concern, there is no electromagnetic interference, it is license free and it can be easily integrated into the existing lighting infrastructure. LEDs are commonly used as transmitters in VLC, while the photodiodes and camera sensors are commonly used VLC receivers. Nowadays, vehicles are equipped with LED lightening due to their long life span, high resistance to vibration, high energy efficiency and better safety performance. Furthermore, VLC receivers are mounted in many vehicles as the front or rear camera for lane tracking and parking purposes. These advantages make VLC a promising solution for ITS applications.

REFERENCES

Abualhoul, M. Y., Marouf, M., Shagdar, O., & Nashashibi, F. (2014). Enhancing the Field of View Limitation of Visible Light Communication-based Platoon. In *Proceedings of the 2014 6th International Symposium on Wireless Vehicular Communications (WiVeC 2014)*, (pp. 1-5). Vancouver, Canada: IEEE. 10.1109/WIVC.2014.6953221

Armstrong, J., Sekercioglu, A. Y., & Neild, A. (2013). Visible Light Positioning: A Road Map for International Standardization. *IEEE Communications Magazine*, 51(12), 68–73. doi:10.1109/MCOM.2013.6685759

Bai, B., Chen, G., Xu, Z., & Fan, Y. (2011). Visible Light Positioning Based on LED Traffic Light and Photodiode, In *Proceedings of the Vehicular Technology Conference (VTC Fall)*, (pp. 1-5). San Francisco, CA: IEEE. 10.1109/VETECF.2011.6092849

Bai, B., Xu, Z., & Fan, Y. (2010). Joint LED Dimming and High Capacity Visible Light Communication by Overlapping PPM. In *Proceedings of the 19th Annual Wireless and Optical Communications Conference (WOCC 2010)*, (pp. 1-5). Shanghai, China. 10.1109/WOCC.2010.5510410

Cahyadi, W. A., Kim, Y.-H., Chung, Y.-H., & Ghassemlooy, Z. (2015). Efficient Road Surface Detection Using Visible Light Communication, In *Proceedings of the 2015 Seventh International Conference on Ubiquitous and Future Networks*, (pp. 61-63). Sapporo, Japan: IEEE. 10.1109/ICUFN.2015.7182498

Cailean, A.-M., Cagneau, B., Chassagne, L., Popa, V., & Dimian, M. A Survey on the Usage of DSRC and VLC in Communication-Based Vehicle Safety Applications. In *Proceedings of the 2014 21st Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, (pp. 69-74). Taipei, Taiwan: IEEE. 10.1109/SCVT.2014.7046710

Cailean, A.-M., Cagneau, B., Chassagne, L., Topsu, S., Alayali, Y., & Dimian, M. (2013). Visible Light Communications Cooperative Architecture for the Intelligent Transportation System, In *Proceedings of the 2013 20th Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, (pp. 1-5). Namur, Belgium: IEEE. 10.1109/SCVT.2013.6736001

Cailean, A.-M., & Dimian, M. (2017). Current Challenges for Visible Light Communications Usage in Vehicle Applications: A Survey. *IEEE Communications Surveys and Tutorials*, 19(4), 2681–2703. doi:10.1109/COMST.2017.2706940

Cheng, L., Viriyasitavat, W., Boban, M., & Tsai, H.-M. (2017). Comparison of Radio Frequency and Visible Light Propagation Channels for Vehicular Communications. *IEEE Access: Practical Innovations, Open Solutions*, 6, 2634–2644. doi:10.1109/ACCESS.2017.2784620

Cui, Z., Wang, C., & Tsai, H.-M. (2014). Characterizing Channel Fading in Vehicular Visible Light Communications with Video Data. In *Proceedings of the Vehicular Networking Conference (VNC)*, (pp. 226-229). Paderborn, Germany: IEEE. 10.1109/VNC.2014.7013353

Deng, Y., Li, L., Wang, H., Liu, Z., Wang, S., & Shen, J. (2015). Visible Light Communication based Collision Avoidance Scheme for Vehicle Lane-changing in Intelligent Transport Systems. In *Proceedings of the 14th International Conference on Optical Communications and Networks (ICOON)*, (pp. 1-3). Nanjing, China. doi: 10.1109/ICOON.2015.7203772

Falcitelli, M., & Pagano, P. (2016). Visible Light Communication for Cooperative ITS. In Alam, M., Ferreira J. & Fonseca J. (Eds.), *Intelligent Transportation Systems, Studies in Systems, Decision and Control* 52, (pp. 19-47). Springer.

Institute of Electrical and Electronics Engineers (IEEE) 802.15.7. (2011). *IEEE Standard for Local and Metropolitan Area Networks—Part 15.7: Short-Range Wireless Optical Communication Using Visible Light*. IEEE.

Intelligent Transport Systems (ITS) Communications Architecture. (2010). *ETSI EN 302 655 v.1.1.1*. Available: https://www.etsi.org/deliver/etsi_en/302600_302699/302665/01.01.01_60/en_302665v010101p.pdf

Khan, L. U. (2017). Visible Light Communication: Applications, Architecture, Standardization and Research Challenges. *Digital Communications and Network*, 3(2), 78–88. doi:10.1016/j.dcan.2016.07.004

Kim, B. W., & Jung, S.-Y. (2016). Vehicle Positioning Scheme using V2V and V2I Visible Light Communications. In *Proceedings of the 2016 83rd Vehicular Technology Conference (VTC Spring)*, (pp. 1-5). Nanjing, China: IEEE. 10.1109/VTCSpring.2016.7504526

Kostic-Ljubisavljevic, A., Radonjic Djogatovic, V., & Mikavica, B. (2017). Visible Light Communication as a Segment of the Intelligent Transportation Systems. In *Proceedings of the Sixth International Conference Transport and Logistics*, (pp. 211-214). Nis, Serbia: Academic Press.

Kumar, N., Lourenco, N., Teraa, D., Alves, L. N., & Aguiar, R. L. (2012). Visible Light Communications in Intelligent Transportation Systems. In *Proceedings of the 2012 Intelligent Vehicles Symposium*, (pp. 748-753). Alcala de Henares, Spain: IEEE. 10.1109/IVS.2012.6232282

Lin, Y., Wang, P., & Ma, M. (2017). Intelligent Transport System (ITS): Concept, Challenge and Opportunity. In *Proceedings of the 2017 3rd International Conference on Big Data Security on Cloud*, (pp. 167-172). Beijing, China: IEEE. doi: 10.1109/BigDataSecurity.2017.50

Liu, C. B., Sadeghi, B., & Knightly, E. W. (2011). Enabling Vehicular Visible Light Communication (V2LC) Networks, In *Proceedings of the Eighth International Workshop on Vehicular Internetworking, VANET'11*, (pp. 41-50). Las Vegas, NV: ACM. 10.1145/2030698.2030705

Luo, P., Ghassemlooy, Z., Le Minh, H., Bentley, E., Burton, A., & Tang, X. (2014). Fundamental Analysis of a Car to Car Visible Light Communication System. In *Proceedings of the 2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP)*, (pp. 1011-1016). Manchester, UK. 10.1109/CSNDSP.2014.6923977

Novak, M., Dobesch, A., Wilfert, O., & Janik, L. (2019). Visible Light Communication Transmitter Position Detection for use in ITS. *Optical Switching and Networking*, 33, 161–168. Advance online publication. doi:10.1016/j.osn.2018.04.002

Park, S.-B., Jung, D. K., Shin, H. S., Hyun, Y.-J., Lee, K., & Oh, Y. J. (2007). Information Broadcasting System Based on Visible Light Signboard. In *Proceedings of the IASTED International Conference on Wireless and Optical Communications*, (pp. 311-313). Montreal, Canada: Academic Press.

Pathak, P. H., Feng, X., Hu, P., & Mohapatra, P. (2015). Visible Light Communication, Networking, and Sensing: A Survey, Potential and Challenges. *IEEE Communications Surveys and Tutorials*, 17(4), 2047–2077. doi:10.1109/COMST.2015.2476474

Petrariu, A. I., Lavric, A., & Coca, E. (2018). VLC for vehicular communications: A multiple input multiple output (MIMO) approach. In *Proceedings of the 2018 International Conference on Development and Application Systems (DAS)*, (pp. 134-137). Suceava, Romania. 10.1109/DAAS.2018.8396085

Qiao, Y., Xia, Z., Hong, X., & Shen, J. (2017). A Visible Light Communication Based Vehicle Collision Avoidance Scheme for Curve Lane Situations. In *Proceedings of the 2017 9th International Conference on Advanced Infocomm Technology (ICAIT)*, (pp. 2141-217). Chengdu, China: IEEE. 10.1109/ICAIT.2017.8388917

Shiu, D.-S., & Kahn, J. (1999). Differential Pulse-Position Modulation for Power-Efficient Optical Communication. *IEEE Transactions on Communications*, 47(8), 1201–1210. doi:10.1109/26.780456

Siddiqi, K., Raza, A. D., & Sheikh Muhammad, S. (2016). Visible Light Communication for V2V Intelligent Transport System. In *Proceedings of the 2016 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom)*, (pp. 1-4). Graz, Austria. doi: 10.1109/COBCOM.2016.7593510

Sjöberg, K., Andres, P., Buburuzan, T., & Brakemeier, A. (2017). Cooperative Intelligent Transport Systems in Europe: Current Deployment Status and Outlook. *IEEE Vehicular Technology Magazine*, 12(2), 89–97. doi:10.1109/MVT.2017.2670018

Sugiyama, H., Haruyama, S., & Nakagawa, M. (2007). Brightness Control Methods for Illumination and Visible-Light Communication Systems. In *Proceedings of the 2007 Third International Conference on Wireless and Mobile Communications (ICWMC'07)*, (pp. 78-84). Guadeloupe, France. 10.1109/ICWMC.2007.26

Sugiyama, H., & Nosu, K. (1989). MPPM: A Method for Improving the Band-Utilization Efficiency in Optical PPM. *Journal of Lightwave Technology*, 7(3), 465–472. doi:10.1109/50.16882

Takai, I., Harada, T., Andoh, M., Yasutomi, K., Kagawa, K., & Kawahito, S. (2014). Optical Vehicle-to-Vehicle Communication System Using LED Transmitter and Camera Receiver. *IEEE Photonics Journal*, 6(5), Article Sequence Number: 7902513. doi:10.1109/JPHOT.2014.2352620

Takai, I., Ito, S., Yasutomi, K., Kagawa, K., Andoh, M., & Kawahito, S. (2013). LED and CMOS Image Sensor Based Optical Wireless Communication System for Automotive Applications. *IEEE Photonics Journal*, 5(5), Article Sequence Number: 6801418. doi:10.1109/JPHOT.2013.2277881

Tram, V. T. B., & Yoo, M. (2018). Vehicle-to-Vehicle Distance Estimation Using a Low-Resolution Camera Based on Visible Light Communications. *IEEE Access: Practical Innovations, Open Solutions*, 6, 4521–4527. doi:10.1109/ACCESS.2018.2793306

Ucar, S., Turan, B., Ergen, S. C., Ozkasap, O., & Ergen, M. (2016). Dimming Support for Visible Light Communication in Intelligent Transportation and Traffic System. In *Proceedings of the 2016 IEEE/IFIP Network Operations and Management Symposium*, (pp. 1193-1196). Istanbul, Turkey: IEEE. 10.1109/NOMS.2016.7502986

Vappangi, S., & Mani, V. V. (2019). Concurrent illumination and communication: A Survey on Visible Light Communication. *Physical Communication*, 33, 90–114. doi:10.1016/j.phycom.2018.12.017

Vucic, J., Kottke, C., Nerreter, S., Habel, K., Buttner, A., Langer, K.-D., & Walewski, J. (2009). 125 Mbit/s over 5 m Wireless Distance by Use of OOK-Modulated Phosphorescent White LEDs. In *Proceedings of the 2009 35th European Conference on Optical Communication*, (pp. 1-2). Vienna, Austria: Academic Press.

Vucic, J., Kottke, C., Nerreter, S., Habel, K., Buttner, A., Langer, K.-D., & Walewski, J. (2010). 230 Mbit/s via a Wireless Visible-Light Link Based on OOK Modulation of Phosphorescent White LEDs. In *Proceedings of the 2010 Conference on Optical Fiber Communication (OFC/NFOEC), collocated National Fiber Optic Engineers Conference*, (pp. 1-3). San Diego, CA: Academic Press.

Wang, N., Liu, C., Lu, Y., & Shen, J. (2017). A Visible Light Communication (VLC) based Intelligent Transportation System for Lorry Fleet. In *Proceedings of the 16th International Conference on Optical Communications and Networks (ICOON)*, (pp. 1-3). Wuzhen, China. 10.1109/ICOON.2017.8121589

You, S. H., Chang, S. H., Lin, H. M., & Tsai, H. M. (2013). Visible Light Communications for Scooter Safety. In *Proceedings of the 11th annual international conference on Mobile systems, applications, and services*, (pp. 509-510). Taipei, Taiwan. doi: 10.1145/2462456.2465707

Yu, S.-H., Shih, O., Tsai, H.-M., Wisitpongphan, N., & Roberts, R. D. (2013). Smart Automotive Lighting for Vehicle Safety. *IEEE Communications Magazine*, 51(12), 50–59. doi:10.1109/MCOM.2013.6685757

Zhuang, Y., Hua, L., Qi, L., Yang, J., Cao, P., Cao, Y., Wu, Y., Thompson, J., & Haas, H. (2018). A Survey of Positioning Systems Using Visible LED Lights. *IEEE Communication Surveys & Tutorials, Early Access*, 1-1. doi:10.1109/COMST.2018.2806558

KEY TERMS AND DEFINITIONS

Infrastructure-to-Vehicle (I2V) Communications: A segment of ITS mainly used for data transmission from roadside infrastructure to vehicles.

Intelligent Transportation Systems (ITS): A set of information and communication technologies applied to the transportation systems with the aim of providing a more efficient, sustainable, and safe transportation system.

Vehicle-to-Infrastructure (V2I) Communications: A segment of ITS mainly used for data transmission from vehicles to the roadside infrastructure.

Vehicle-to-Vehicle (V2V) Communications: A segment of ITS mainly used for data transmission between vehicles on the roadside.

Visible Light Communication (VLC): An optical wireless communication technology using visible light as a communication medium.

VLC Receiver: Device that receives information modulated in the visible light spectrum. Photodiodes and camera sensors are commonly used.

VLC Transmitter: Device that transmits information using visible light spectrum. LEDs are commonly used.

Section 8

Digital Media

Antecedents and Consequences of the Growing Popularity of Digital Consumption

Ritu Narang

University of Lucknow, India

Sonal Tiwari

University of Lucknow, India

INTRODUCTION

There is something addictive and enthralling about the world where everything is digital (Hoffman, Novak and Stein, 2012). Splurging a large amount of one's time online, executing more than one task simultaneously and acquainting with new technologies easily are all testimonies to the increasing digital nature of everyone's daily life. Hoffman et al. (2012) further stated that new media has become an inevitable part of consumer's daily life and routine. It is influencing, altering and transforming the way they communicate, work, play, read, write, watch, listen, retain, explore, browse, shop, collaborate, produce and create. Each of these actions in the post-digital era was quite different when compared to the pre-digital era. New media has completely revamped people's everyday lives. For majority of people now, logging in is the foremost thing in the morning even before getting up (Hoffman et al., 2012). Stone (2009) commented in his study that those days are gone when newspaper was the only source of distraction while having first meal of the day with family. Majority of people now have replaced newspaper with mobile phones, tablets, laptops and other technological devices during breakfast with the family which was once known as the family time. People may spend the day on their smartphones using various applications either to browse restaurants, book movie tickets, view television shows online, listen to music, shop online, play games or to keep track of their daily exercise (Hoffman et al, 2012). Technology is present and apparent in every aspect of people's daily lives. Now, they have become addicted and dependent on new media to such an extent that they may become restless and feel panic if their internet connection is down for even a short period of time (Hoffman et al, 2012).

Thus, seeing the growth rate in the new technology and consumption of the internet, the current chapter tries to explore the reasons for the growth of digital consumption and consequences due to its over consumption. Till date, majority of studies have focussed on the factors which lead to internet addiction in general (Griffiths, 1995; Beard, 2005; Young, 1998; Caplan, 2005; Kim, LaRose & Peng, 2006). However, aspects of digital consumption like online television, online games, etc. have not been explored much. Therefore, the chapter tries to bridge the gap by taking into account 'internet television', a significant area under digital consumption. The present chapter seeks to draw attention on the following aspects of digital consumption:

- It will elaborate upon the concept of digital consumption and its growing popularity.
- It will focus on the meaning of internet television and reasons behind its growth.
- It will throw some light on the behavioural and physiological consequences of internet television consumption.

DOI: 10.4018/978-1-7998-3479-3.ch073

- It will include a conceptual model and its implications followed by some recommendations, insights about future scope of the study and conclusion.

BACKGROUND

Digital Consumption: It's Growing Popularity

With proliferation of digital devices and increasing internet penetration, today's consumers are more connected than ever. The increase in internet penetration and information and communication technology has given rise to new distribution channels, making online presence a requisite condition for marketers to extend their geographical boundaries (Narang & Trivedi, 2016). Earlier, content was available only through few mediums such as print, radio and conventional television but now it is also available through numerous multiple connected devices. This has revolutionised the media world and obscured traditional media definitions (Nielsen, 2014). The internet is widely used these days, with more than 3.75 billion people online today (Digital in report, 2017). Internet use continues to grow all over the world, with global users rising at the rate of 8 per cent per year. Today, nearly two-third of the world's population use smartphone and majority of the world's web traffic comes from smartphones. There are around 4.92 billion people across the globe that uses mobile phones which is equal to 66 per cent penetration (Digital in report, 2017).

Billions of people throughout the length of the day are continuously engrossed in the consumption of the internet, social media, mobile applications and other digital communication technologies (Stephen, 2015). This digital nature of people can be observed by looking at the time they are spending on the internet and the way they are embracing new technology (Belk and Liams, 2013). For instance, the internet usage rate among Americans is close to 87 per cent (Pew Research Center, 2015). Majority of people now have smartphones, high-definition televisions, computers, laptops, wi-fi networks due to which they are spending on an average of 60 hours per week on the consumption of content across various platforms or screens (Nielsen, 2014). In India, around 180 million people view online videos through YouTube on mobile alone (Chaturvedi, 2017). According to State of digital downloads research report (2017), over 62 per cent of European respondents prefer to download or watch TV shows and movies online. France has witnessed the highest streaming rate i.e. 49 per cent when compared to the UK and Germany who have witnessed 42 and 40 per cent respectively. Around 37 per cent of German consumers, 36 per cent of UK consumers and 30 per cent of French consumers prefer to download books and magazines. Similarly, 34 per cent of Germany and UK consumers choose to download music, on the other hand 56 per cent of French consumers more likely to stream music. In fact, social media has witnessed tremendous growth worldwide; more than 2 billion people are using it among which Facebook alone has 1 billion active users per day (Stephen, 2015). Nowadays social media usage has become an inevitable practice in consumers' daily lives. The number of social media users across the globe grew by more than 20 per cent over the past 12 months, with well over one-third of the world's population now using social media every month. Mobile social media use is growing even faster, with global monthly active user numbers up 30 per cent year-on-year to reach more than 2.5 billion (Digital in Report, 2017). Around 64 per cent of total social media users stated that they use these sites at least once a day through their computers and 47 per cent of smartphone owners stated that they spend time on these sites every day (Nielsen, 2014).

Initially, only the developed countries indulged in digital consumption but now even the developing countries are consuming it. As per the KPMG International Report (2013), people of China, Brazil and

Singapore are leading consumers of digital media and are consuming internet at a substantial pace. In fact, consumers of these countries are also ready to pay for the consumption of digital media. 78 per cent of Chinese consumers own a smartphone which showcases higher penetration as compared to the US, UK, Australia and Germany. With 140 million internet users Brazil has become the largest internet market in Latin America and fourth largest across the globe (Dossier, 2016). In South Korea, which is one of the most digitally advanced countries across the world, government reckoned that around 2.55 million people are addicted to their smartphones by consuming it for more than 8 hours per day (Yeomans, 2013). In India, there are around 3 lakh app developers and the country has already become second largest android developer community in the world (Deloitte, 2015). In 2014, smartphones users went beyond 2 billion and are forecasted to touch 4.6 billion mark by 2019 (Deloitte, 2015). The EY (2016) anticipates Indian smartphone market to reach 520mn by 2020, which would make India emerge as one of the largest smart phone economies in the world. With the emergence of better networks, coverage and advanced technologies such as 3G and 4G, there is an immense rise in the data consumption across the globe and it is quite evident in India with the launch of Jio by Reliance in the year 2016. Highest internet penetration is in countries like United Arab Emirates, Iceland, Norway, Luxembourg, Denmark, Bermuda, Andorra, Netherlands, Bahrain and Japan with 99 per cent, 98 per cent, 97 per cent, 96 per cent, 95 per cent and 93 per cent respectively (Digital in report, 2017).

Meaning of Internet Television

This section seeks to explore the meaning of Internet Television and reasons behind its growth. Internet television has different names such as personal TV, web TV, interactive TV, enhanced TV and so on (Gerbarg and Noam, 2004). According to Noll, (2004) internet television also known as online television is nothing but a traditional television which is now available on the internet. It is a video service that is available online which operates through website streaming in order to offer various television shows or videos (Schechner & Stewart, 2012). It is a streaming media through which people do not have to wait to download the video; instead, they can watch the television shows in real time as they are sent in a continuous stream of data and can be played as they are uploaded on the internet. Internet television has variously been defined as:

“Internet TV is an online video service that uses website streaming to offer TV programs or videos”
(Li, 2013, p.6)

“Online programming that makes media content available through a computer screen, tablet or speaker. It has the ability to displace or substantially supplement the use of non-computer media content”. (Ferguson, 2012, p. 143)

“Internet television is an internet subscription service that allows audiences to watch TV shows, movies, or videos via any digital devices via internet streaming”. (Schechner and Stewart, 2012, p. 9)

In this chapter, internet television refers to all those content which are produced primarily for viewing on the television set via cable or satellite, which is later made accessible online via various applications; content produced primarily for online distribution such as various web-series and user-generated content which can be produced by anybody and made available through YouTube.

REASONS FOR THE GROWTH OF INTERNET TELEVISION CONSUMPTION

There are several studies which have identified television viewing motivations and gratifications. Katz, Blumler and Gurevitch (1973) classify media consumption needs into five categories- Cognitive needs (understanding, knowledge and Information), Personal integrative needs (status, stability, confidence and credibility), Affective needs (emotional pleasurable or aesthetic experience), Social integrative needs (strengthening contacts with friends, family, etc.) and Tension release needs (diversion and escape). According to Greenberg (1974), there are seven motives behind television viewing- habit, relaxation, passing time, arousal, escape, companionship and learning. Palmgreen and Rayburn (1979) mention seven television viewing gratifications such as passing time, entertainment, relaxation, learning, forgetting, communication utility and companionship. Rubin (1981) determines nine television viewing motivations that include escape, companionship, passing time, habit, arousal, relaxation, entertainment, social interaction and information. Several studies have identified socio cognitive motives of television viewing. These motives were classified into two categories: positive and negative. Positive motives comprise of social engagement and enjoyment while, negative motives include boredom, escape from reality and finding solitude and solace among fictional TV characters (Farber, Khavari and Douglass, 1980; Cooper, 1994; Cooper et al. 1995; Lang et al. 2012 and Bailey and Baillie, 2013).

Ferguson and Perse (2000) postulate that motives like entertainment, relaxation and time pass are the same for both television and internet consumption and thus the internet may be functionally equivalent to television and can be treated as its substitute. Similarly, there are a few studies that also indicate that the motives for viewing television and internet television are identical such as information, escape, entertainment and social interaction (Ambra and Rice, 2001; Kaye, 1998; Lin 2001). On the contrary Li (2013) posits that motives of both the mediums are quite different. According to him, viewers attain three types of gratifications- content, process and social. Content gratification encompasses time pass, information and entertainment. Convenience and ease of use come under process gratification. Social gratification incorporates companionship and social interaction. In the opinion of Li (2013) process gratification is derived from the consumption of internet television and content gratification is obtained from television consumption. Harrell (2000) argues that internet users choose mass media in order to satisfy their information, companionship, escapism, communication and entertainment needs. According to Shin (2007), motives of internet television consumption can be divided into two: extrinsic and intrinsic. Extrinsic motives include highly interactive services and interoperable applications and intrinsic motives comprise of rich content, high quality videos and value added services.

Seeing the growth rate in the consumption of various digital platforms in various countries one becomes curious to explore the reasons behind this phenomenon. Some of the reasons are as follows:

Content

With the evolution of so many digital platforms, people are now able to consume different formats of media online such as movies, news, short form content, social interaction, gaming and education (EY, 2016). There is an increase in the inclination towards online content platforms. Multiple players in this sector are now offering personalised content in the form of web-series, videos and other digital programming options in order to generate the interest of their viewers and make their eyeballs glued to the digital screens. Viewers are now getting far more individualized experience due to the internet. Since so many shows are available online, computer screen has easily substituted television screen (Marshall, 2009).

Earlier, content was created in multiples of 30 minutes but with the emergence of on-demand viewing, people are no more interested in spending an hour or half-an-hour in front of their television screens. Nowadays, consumers, especially the younger lot are viewing more of short-form content whose average length is less than 20 minutes. According to the EY Report (2016, p 9), “62 per cent of the content consumed on YouTube is short-form content.” Due to this trend, content producers have now started to focus more on the production of short-form content. From this, it can be induced that today’s consumers are fascinated more by the story line and not its length.

Convenience

The option to watch several videos and TV shows anytime and anywhere gives rise to consumption of Internet Television. It means now viewers do not have to make themselves available at the time of broadcast in order to watch the latest episode of a favourite show. Internet viewers do not have to wait for the show to be aired instead they can watch them as per their own convenience. Due to this trend, more and more networks are now making their shows available online (Melton, 2013). Consumers or Audiences who have an option to skip, can catch their favourite shows on various digital platforms in case they miss it live on television. Usually, evening time is considered as a prime time for watching a television show but now they have an option to view a range of shows at any time during the day or night according to their convenience (Pena, 2015). According to a research conducted by comScore Inc., (2014) about 56 per cent of the US internet users stated that they prefer to watch television online because it gives them an opportunity to watch their favourite shows as per their own schedule. While 50 per cent of the users stated that they view television on internet because it is just more convenient.

With the advent of several sites and mobile applications such as Netflix, Amazon Prime, Hulu and many more has given a platform through which ample of television shows and videos can be accessed rapidly and inexpensively (Lee, 2013). The ever-increasing popularity of smartphones, tablets and laptops has led to a phenomenon known as “second screening” which means viewers are using internet connected gadgets in order to avail additional content be it watching any show or using social media handles. The best thing about viewing television on these internet-connected devices is that they are handy and mobile i.e. consumers can take these devices along with them irrespective of where they are heading. Viewers who prefer to watch online television on these devices are not restricted to view it only where their television is placed at home. They can view it at the gym, lunch breaks, while waiting in a queue or anywhere else (Melton, 2013).

According to a study conducted by the Online Publishers Association, (2004) two kind of lifestyles exist among the consumers who are aged between 18-34 years. These lifestyles can be categorised as “unsettled” and “newly settled”. Consumers who all are aged 18-24 years are likely to fall in the first category i.e. unsettled. They are the ones who do not have any predictable schedule or routine, for instance, a college student may choose to go to the gym one day while other day he might choose to socialize with friends. In general, these unsettled consumers do not follow a set schedule or a prescribed routine as compared to the consumers who are newly settled. Thus, as a result, convenience and accessibility play a key role in swaying media use. The study further stated that various media serve different functions like television is treated as a source of escape and entertainment while radio is treated as a source of companionship and connectedness and web or the internet is a source of information and described as convenient and accessible.

Binge Watching

The tendency of viewers to be hooked to online television is known as ‘binge watching’. It is also known as ‘marathon viewing’ and is a recent behaviour which is found among the consumers these days. It means “viewing multiple episodes of a particular show or series over a concentrated period of time” (Pena, 2015). This recent television behaviour allows consumers to view their favourite show all at once at their convenience (Sodano, 2012). Horsey (2013), stated that there is no such thing as missing a show in this new era of television. VCRs which were used earlier went digital and turned into DVR which made it possible for consumers to record their favourite show and watch it later on (Chmielewski, 2013). Soon after, with the further advancement of digital services, such as, Netflix, Amazon Prime, Hulu and video on demand, the habit to consume all the episodes of a show at one go has been elevated (Umstead, 2014). A consumer can indulge into binge watching whenever and wherever he or she wants. He or she can spend his or her summer break by watching movies without a break or the complete season of his or her favourite show while fighting the flu (Pena, 2015).

Pena, (2015) found in a survey that people do not indulge in binge watching intentionally. Around 71 per cent of binge watchers claimed that initially they did not intend to watch more than one or two episodes but once they started watching it, they ended up watching more episodes. A 2013 survey conducted by MarketCast unveiled that around 67 per cent of American television viewers who are aged between 13-49 years old, are engrossed in binge watching. A survey done by TiVo, (2014) revealed that 91 per cent of its subscribers see binge viewing as a common viewing behaviour. These figures are likely to increase more with the launch of various technology devices and television content which are now easily available to consumers. The services such as Hulu, Amazon Prime, Netflix, HBO Go, HotStar and many more are constantly devoted to making binge viewing convenient to consumers and remunerative to business (Pena, 2015).

Control

Gone are the days when viewership norms were based on mass appeal, finite shows on set schedules, few genres and limited access and control by the consumers. The television industry was dominated by very few networks, offering unique shows, uniform programming and availability, which compelled audiences to settle for the same, the only option they had was either to watch it in daytime or prime-time (Lotz, 2007). Networks competed with each other for securing prime time slots and kept the viewers on the hook by ending each episode with suspense afford to miss (Barker, 2011). It was the time when television was known as “appointment television” where audiences used to make themselves available by altering their schedules so that they could watch their favourite shows at the scheduled time (Pena, 2015).

Appointment television basically refers to the popular shows that people make a point to remember to watch. Now with the expansion of technology, television has moved to the internet and does not remain time-constricted, thus lifespan of appointment television is now uncertain. Internet television has challenged the perception that viewers would still permit the cable channels and networks to tell them ‘what’ and ‘when’ to watch (Zurawik, 2013). Steinkamp, (2010) conducted a survey and stated that out of 136 respondents, 70 vehemently agreed and 41 agreed that Internet Television enables them “to choose when to watch my favourite TV program”.

Internet has devastated the television industry, but not the television. Television still is treated as a platform which has a power to bring people together by making them watch their favourite shows and getting them invest in its storyline and characters. Viewers still are longing to know what happens to

their favourite character in the show. The only thing which has changed now is that they want to know it at a time that suits them and not the network (Montpetit, 2014).

Skip Commercials

Skipping commercials is one of a few added perks which viewers receive while watching television online. Viewers get an option to skip commercials and enjoy watching their favourite television shows online without any disturbance. Consumers always try to find ways to avoid commercials or advertisements. This is why most of the consumers end up purchasing ad-free subscription services some of which are Amazon Prime and Netflix so that they can zip through the advertisements (James and Battaglio, 2017). Steinkamp (2010), conducted a survey in which one of the respondents stated that it's quite annoying whenever commercials pop up while he/she is watching a show. At least with online television they get an option to skip commercials and even if they won't get this option then in that case they only have to view these ads for 30 seconds max. Another respondent stated that nowadays, television commercials are becoming longer and more obnoxious and he was of the opinion that shows with fewer commercials were better.

Consequences of Internet Television Consumption

This section throws some light on the consequences or effects of post digital consumption on its users. In the present scenario, smartphones have become our constant companion: WhatsApp and Facebook for connecting with friends, Netflix for watching movies and shows, etc. Some consumers use their smartphones so many times in a day that soon they start generating this feeling that they cannot live without it. New media when consumed excessively may lead to addiction. It was Marks, (1990) who put forth the concept of media addiction. According to him, there is an analogy between media addiction and psychiatric conditions with dependency or compulsive qualities or a sort of obsession, and the criteria to diagnose those conditions were accordingly adjusted to media consumption behaviour.

An American psychologist, Goldberg, (1996) evolved a new concept known as internet addiction. According to him, internet addiction is a process of reducing stress, but excessive addiction might result in weakening the mental and physical function of a user's career, social nature and relationship with family. Few studies (Young, 1996; Greenfield & Davis, 2002) have compared media addiction to the term 'impulse control disorder' and have clarified that it does include an inebriant characteristic. Kandell, (1998) defined media addiction as a psychological habituation of media and according to Griffiths, (2000) it is a technological addiction and a subgroup of behavioural addiction. Many researchers have identified different terms for media addiction such as 'unregulated media use' (LaRose, Lin & Eastin, 2003), 'media dependence' (Yuen & Lavin, 2004), 'compulsive media use' (Meerkerk, Eijnden & Garretsen, 2006), 'problematic media use' (Caplan, 2002, 2010; Davis, Flett & Avibesser, 2002) and 'excessive media use' (Suhail & Bargees, 2006). According to Kalus (2016), addiction develops in five stages:

- **Trying:** Addiction begins with the trying of new things. For instance, playing a new game, watching a new show, etc.
- **Enjoyment:** When a person is delighted by using a new media application.
- **Habituation:** When a person inculcate a habit of using new media application.
- **Misuse:** When media consumption goes beyond normal consumption.
- **Addiction:** When a person starts neglecting his or her work, family, friends, daily hobbies in order to use this application.

Needless to say, consumers have to witness certain consequences or implications whenever they consume anything in excess and internet is no exception. According to Xu, *et. al.* (2018), health does not merely mean absence of disease but it comprises of physical, mental and social well-being. Physical health allows individuals to perform their routine activities without any restrictions. Mental health deals with a state of well-being in which the individual realises his or her own abilities and can work productively and fruitfully. Social health refers to an individual's ability to manage in social situations and behave accordingly. In this chapter, consequences have been divided into two- physiological and behavioural. Physiological consequences comprise of physical health and behavioural consequences deal with mental and social health. The consequences are as follows:

Physiological Consequences

Wang et al, (2015) stated that excessive use of smartphones lead to depression, sleep disorder and psychological discomfort. According to Young, (1998) some physical problems were faced by the people due to over consumption of internet such as poor immunity system, fatigue and lack of sleep. Brenner, (1997) conducted an online survey on the internet usage and came to the conclusion that people who spend more time on the internet are more likely to face several problems such as missing meals, lack of time management, lack of sleep, etc. People who are internet addicted tend to suffer from physical symptoms such as insomnia, fatigue, exhaustion, poor eyesight, appetite problems, weight loss, weight gain (due to physical inactivity), agitation, backache, neck pain and arms pain (Lan and Lee, 2013).

Behavioural Consequences

The use of several personal devices such as mobile phones, laptops, tablets, etc. to view various shows, movies and other videos lead to “personal escapism” which means viewing content individually and not with the family (EY, 2016). This trend might change the type of content consumed by the viewers. Viewers come across to many shows, movies or games that consist of violence, which in turn can influence their behaviour. According to Funk (2004), Violence in entertainment media acts as a contributor to violent and aggressive behaviour in real life. Farrell and Bruce, (1997) conducted a survey on students in order to ascertain the impact of violence exposure on the level of their aggression and distress. They concluded that there was a significant relation between violence exposure and aggression. Farrell and Bruce suggested that children who view things which consist of violence become desensitized to violence, which increases aggression. Media presentations of justified violence may change a person's belief that violent behaviour is undesirable and wrong. This change in one's belief might promote pro violence attitudes (Funk, 2004). Anderson and Dill, (2000) differentiate between short-term and long-term effects of violence shown in video games on consumers. Short-term effects of video game are concerned with the present internal state of a person. Whenever a person is exposed to violent media, it gives rise to aggressive thoughts, negative feelings or increase in arousal. Long-term effects may result “in the development, over-learning and reinforcement of aggression-related knowledge structures” (p-9). These knowledge structures consist of aggressive action against others; conjecture that other people will behave aggressively, positive attitude towards use of violence and the belief that violent solutions which are derived through a violent action are acceptable and effective. According to Gentile, (2000) the more the students are exposed to violent media, the more they indulge in physical fights and arguments with teachers.

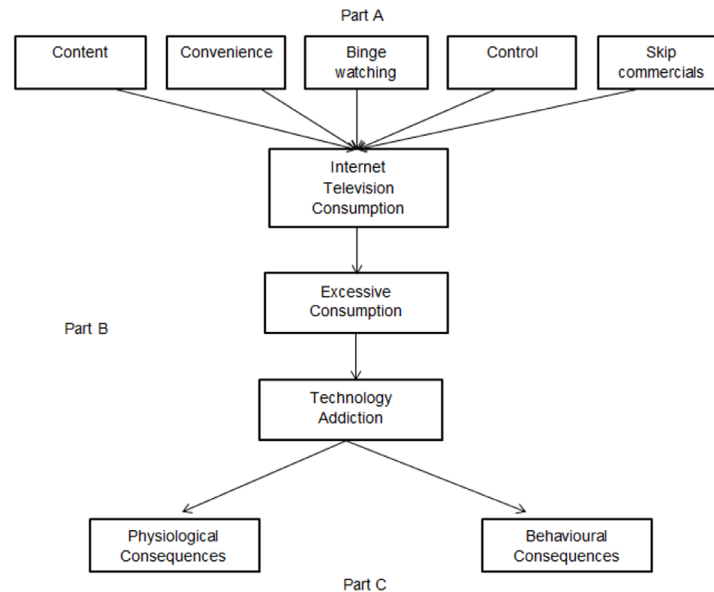
Some internet users stated that they are witnessing quite serious consequences such as becoming socially isolated or getting into trouble with their boss (Brenner, 1997). According to Kubey & Csikszentmihalyi (2002), people who are addicted to television are less likely to engage in sports and other community activities. Few researchers argued that television addiction hampers one's ability to function at work and home (Meerlo, 1954; Sjoberg & Magneberg, 2007). People tend to give up their social, recreational and occupational activities for instance, one may start ignoring their jobs or family duties or prosocial hobbies (McIlwraith, 1998). According to Cao & Su (2007), people with an internet addiction tend to experience greater mental illness than those not addicted. Bandura, (1986, 1991, p. 232) defined deficient self-regulation of internet use as *"a state in which conscious self-control is relatively diminished"* It particularly means when a person fails to monitor his consumption or usage behaviour. Consequently, deficient self-regulation of one's internet use may impact one's professional life and personal relationships adversely (Caplan, 2010). According to Harris and Williams, (1985) students who invest more time and money on playing video games tend to score poor grades in school. The content of the games and shows or movies which students play and watch respectively may hamper their school performances. For instance, if a student is investing more of his time on educational games and shows then he would probably be less likely to show a poor school performance. According to Lieberman et al. (1988) children who use computers in order to play more games, tend to perform more poorly in school. On the other hand, children who use computers for school work, perform better. Gentile, (2004) stated that if children spend more time on playing games rather than on educational and social activities, then, in that case, their grades are affected negatively. Following this logic, it could be inferred that electronic media can impact a person's learning and social behaviour by replacing other activities such as family interaction, reading and social life with peers and friends (Huston, et al., 1992). Young, (1998) classified the problems associated with internet addiction under five categories: relationship, occupational, academic, physical and financial. According to him, excessive use of the internet can have an adverse impact on friendships, parent-child relationships, and husband-wife relationships because they are likely to spend less time with their family and friends. Using internet for personal work at the working hours can have a negative impact on one's productivity which in turn, can result in losing job. Students face academic related problems, when at the time of study they indulge in chatting with friends and exploring immaterial sites.

Shatton (1991) is the one who explored the positive effects of computer addiction. He stated that individuals who overuse computers are likely to gain some benefits from it like people end up exploring improved job offers due to elevation in their analytical, reasoning and technical skills. This knowledge has a positive impact on their confidence and self-esteem. People are able to communicate in an effective manner because of the knowledge they have gained and it has also helped them to expand their friends circle and have an impressive social life. People have gained respect through their expertise in technology and many dependants derived pleasure by giving advice to those ones who were not experienced/well-versed with the new technology Usage of computers has resulted into gaining a sense of achievement and satisfaction which in turn has helped people to relieve them from depression, stress and boredom (Shatton, 1991).

CONCEPTUAL MODEL

Based on the extensive literature review and the observations gained therein, this chapter has proposed a conceptual model to extend knowledge about antecedents and consequences of internet television

Figure 1. Antecedents and Consequences of Internet television consumption



consumption. The proposed conceptual model provides insight into what motivates viewers to consume internet television and consequences that they face due to its overconsumption. Figure 1 illustrates the three main parts namely, A, B and C of the model. Part A deals with the antecedents of internet television consumption, Part B talks about how excessive consumption leads to technology addiction and Part C explains the consequences of technology addiction.

Part A talks about antecedents of internet television, namely content, convenience, control, binge watching and skip commercials.

Content: With content, the researchers mean the storyline, time duration, picturisation and number of episodes of the shows available on several video streaming portals. The demand of internet television increases due to the ‘quality of content’ which these video streaming portals are offering to its viewers (Marshal, 2009 and EY Report, 2016). People get shows of different genres on these portals. Unlike television, there is an unlimited pool of shows available on the internet among which viewers can watch any show of their choice. Furthermore, the quality of actors, story and picturisation of internet shows are far better than the shows available on television. These shows have limited number of episodes and are of shorter duration unlike shows which are available on conventional television. Thus, the availability of good quality content provides further impetus to growth of internet television.

Convenience: Convenience means something that makes things easier, comfortable and more suitable for people (Narang & Trivedi, 2016). One of the reasons for increase in internet television consumption might be its convenient nature. Due to online television, consumers get flexibility of watching what they want at their convenience. People can view internet television anytime and anywhere. Since internet television can be watched on smart phones and laptops, it becomes quite handy and mobile for viewers to see their favourite shows and thus, viewers now do not have to sit at one place on a scheduled time in order to watch television.

Binge watching: Binge watching means viewing episodes of a show one after the other. On the video streaming portals, viewers have an opportunity to watch their favourite shows in one go since all the episodes are available on it and with the availability of internet television, people do not have to wait for the next episode to be aired on television. According to Pena (2015), a consumer can indulge in binge watching whenever and wherever he or she wants due to the availability of all the episodes of his or her favourite shows. Thus, the simultaneous availability of all the episodes on internet might lead to consumption of shows on video streaming portals.

Control: Control refers to an act of ruling something or someone. One of the major reasons viewers might prefer to watch shows online is the power to enjoy control. Today's youth want to be entertained with content made for them to be consumed on devices, time and place of their choice (Bansal, 2018). With the emergence of several video streaming portals, people are not controlled by the cable channels and networks anymore. It is now the consumers who decide 'what' to watch, 'when' to watch, 'where' to watch and on 'which' device to watch their favourite shows, thereby giving absolute dominance in the hands of the viewers.

Skip commercials: Skip commercials refer to a feature that makes it possible for viewers to watch shows without advertisements or commercials. This feature helps in making shows commercial free. People might also choose to watch internet television because shows which are available on the internet are either commercial free or commercials are of maximum 30 seconds which they can easily skip. People do not like to get disturbed while watching something of their interest. These commercials act as a disturbance that could be a reason why people prefer watching shows on the internet and not on conventional television.

Part B talks about excessive use of internet television which might result in addiction on the part of the viewers. Addiction develops in five stages: trying, enjoyment, habituation, misuse and addiction (Kalus, 2016). Initially, people try video streaming portals because of the antecedents mentioned above. Soon they start enjoying the same because of the perks which they get through these shows such as quality of content, power to control, its convenient nature, availability of all the episodes and the option to skip commercials. Then, it becomes their habit and they start consuming it beyond a certain limit and once they over-consume it, they become addicted to it. As stated by Xue, Dong, Luo, Mo, Dong, Zhang and Liang (2018), excessive use of the services can lead to technology addiction.

Users who are technology addicted might experience negative consequences such as physiological and behavioural as explained in Part C of the model. Physiological consequences such as fatigue, lack of sleep, missing meals, lack of time management (Brenner, 1997; Young, 1998; Wang *et. al.* 2015) deal with the physical health of an individual i.e., all internal and external organs and body parts of a person are functioning properly as they are supposed to function. When a person becomes addicted to these video streaming portals and watches shows at night for a longer period of time, they tend to become sleep deprived. Due to lack of proper sleep they are likely to feel drowsy during the day. Behavioural consequences like personal escapism, aggression, violence, social-isolation, poor performance, relationship and occupation related problems, etc. (Harris and Williams, 1985; Brenner, 1997; Young, 1998; Gentile, 2000) deal with the social and mental health of an individual i.e., how a person is behaving in social situations and is he/she able to cope with the normal stresses of life. However, internet television has not only revamped consumers' way of viewing television but also provided certain benefits to its consumers such as gaining expertise in technology, improved communication skills, impressive social life, increase in confidence and self-esteem and relieve from depression, stress and boredom (Shatton, 1991).

IMPLICATIONS OF THE MODEL AND RECOMMENDATIONS

Video streaming portals form an important part of our everyday lives now. Consumers find themselves discovering new shows and binge watching them till they have finally completed a “season”-sometimes in a single night. It rescues them from boredom, exposes them to a new kind of culture and sometimes acts as a topic over which people bond. The shows available online are highly engaging and leave people craving to watch one episode after the other.

No doubt, one cannot ignore the popularity of this new media because they are the source of entertainment. However, this model seeks to bring attention towards negative consequences that might arise due to its excessive consumption. As it is rightly said that *‘there is no rose without a thorn’* similarly, its over-consumption might have some serious impact on viewers. Late night binge watching leads to lack of sleep which might affect an individual’s capacity to work, productivity and concentration level at work, school or colleges. The excitement of “what happens next” in the series might promote people to finish their work as soon as possible without actually paying attention to its quality. Sometimes, they may even procrastinate. Students may find themselves compromising on their studies just to finish a ‘season’ before their peers do.

People who prefer watching shows at night and for longer period are likely to experience mood swings and lower tolerance levels due to lack of proper sleep hence, might become more irritated and aggressive which will further disrupt their personal and professional lives which will ultimately affect the society as well as the economy of the nation. Movies and TV shows have always had an impact on people’s lifestyles-the way they dress up, talk, the kind of jargons they use, the kind of romantic involvements they expect, etc. (Bleakley & Ellithorpe, 2016). There is not a single sphere of life which has not been influenced by the kind of movies and shows people watch. Many researchers have pointed out how early exposure to violent TV shows results in higher levels of anxiety and aggression in children (Farrell & Bruce, 1997, Anderson & Dill, 2000 & Funk, 2004). These studies may have focussed more on children but same might be applied to youngsters also. Youth tend to look for role models in their favourite shows and replicate their actions. If these characters are too violent, aggressive and villainous, they are likely to absent mindedly copy their acts in their peer groups. Their young minds are likely to get negatively influenced and are unable to distinguish right from wrong. Some of the content is so violent that it can render even the adults sleepless.

The model emphasises the need to pay attention to the negative consequences of watching shows on video streaming portals. Family, teachers, media, government, and society can play an important role to reduce the impact of over-consumption on viewers. Government should encourage the development of certain applications which can monitor viewers’ screen time or applications that can lock these video streaming portals for a specified time once it exceeds a certain limit. However, maintaining a disciplined lifestyle and understanding the harmful impact of spending a lot of time watching can be instrumental in controlling this habit. It is also suggested that parents and teachers should monitor the content their children or students are exposed to, also the censor boards should assume a more responsible role in order to see whether the content going on air is refined and not too disturbingly violent in nature. Therefore, only if media is consumed in a proper manner and within limits, it can have positive implications else it could lead to negative implications on an individual as well as the society.

FUTURE RESEARCH DIRECTIONS

Future research can be conducted to empirically test the reasons behind growing popularity of internet television. This chapter has not explored certain other aspects and features associated with internet television such as buffering speed, downloading time, prices, etc. Further studies can also throw some light on the consequences other than the ones which are mentioned in this chapter such as psychological consequences, financial consequences, etc.

CONCLUSION

The objective of the chapter was to explore the following aspects of digital consumption: 1) The concept of digital consumption and its growing popularity. 2) The meaning of internet television and reasons behind its growth. 3) Consequences of internet television consumption. 4) A conceptual model based on antecedents and consequences of internet television consumption. Viewers consume internet television due to its convenient nature, quality of content, control, opportunity to binge watch and option to skip commercials. Thus, the growth rate witnessed by internet television across the globe promises to transform the way consumers or viewers used to watch television in the pre-digital era. The new media technology has not only induced the lifestyle patterns of consumers but also has created a modern culture by swaying the way people communicate, learn, behave and perceive themselves and the world in which they are living. However, if it consumed excessively then the consumers tend to suffer from certain physiological and behavioural consequences.

REFERENCES

- Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology*, 78(4), 772–790. doi:10.1037/0022-3514.78.4.772
- Bandura, A., & National Inst of Mental Health. (1986). *Prentice-Hall series in social learning theory. Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Bandura, A. (1991). Self-efficacy mechanism in physiological activation and health-promoting behavior. In J. Madden IV, (Ed.), *Neurobiology of learning, emotion and affect* (pp. 229–270). Raven.
- Bansal, S. (2018). Are young Indians going from TV to online content. *livemint*. Retrieved January 10, 2019 from <https://www.livemint.com/Home-Page/NdJMoFReIGKgo3uFw5dElK/Are-young-Indians-going-from-TV-to-online-content.html>
- Barker, C. (2011). Television and New Media: Must-click TV, by Jennifer Gillian. *Popular Communication*, 9(3), 232–233. doi:10.1080/15405702.2011.587170
- Beard, W. K. (2005). Internet addiction: A review of current assessment techniques and potential assessment questions. *Cyberpsychology & Behavior*, 8(1), 7–13. doi:10.1089/cpb.2005.8.7

Belk, R. & Liasas, R. (2012). The Digital Consumer. *The Routledge Companion to Digital Consumption* Routledge, 28-38.

Bleakley, A., & Ellithorpe, M. (2016). How are adolescents affected by what they watch on TV. *Annenberg School for Communication*. Retrieved April 17, 2019, from <https://www.asc.upenn.edu/news-events/news/how-are-adolescents-affected-what-they-watch-tv>

Brenner, V. (1997). Psychology of computer use: XLVII. Parameters of Internet use, abuse and addiction: the first 90 days of the Internet Usage Survey. *Psychological Reports*, 80(3), 879–882. doi:10.2466/pr0.1997.80.3.879

Caplan, S. E. (2005). A social skill account of problematic internet use. *Journal of Communication*, 55(4), 721–736. doi:10.1111/j.1460-2466.2005.tb03019.x

Caplan, S. E. (2010). Theory and measurement of generalized problematic Internet use: A two-step approach. *Computers in Human Behavior*, 26(5), 1089–1097. doi:10.1016/j.chb.2010.03.012

Carporelli, L. (2004). Convenience and Accessibility Key Factors in Media Use of 18 to 34 Year-Olds, According to New Online Publishers Association Report. *Business Wire*. Retrieved August 28, 2018, from <https://www.businesswire.com/news/home/20040420005990/en/Convenience-Accessibility-Key-Factors-Media-18-34>

Chaturvedi, A. (2017). YouTube celebrates growth of online creators in India. *The Economic Times*. Retrieved July 18, 2018, from <https://economictimes.indiatimes.com/tech/internet/youtube-celebrates-growth-of-online-creators-in-india/articleshow/57798935.cms>

Chmielewski, D. C. (2013). Binge-viewing is transforming the television experience. *LA Times.com*. Retrieved June 26, 2018, from <http://articles.latimes.com/2013/feb/01/entertainment/la-et-ct-binge-viewing-2013201/2>

comScore. (2014). Why watch online TV? It makes life easier. *eMarketer*. Retrieved from <https://www.emarketer.com/Article/Why-Watch-Online-TV-Makes-Life-Easier/1011388>

Davis, R. A., Flett, G. L., & Besser, A. (2002). Validation of a new scale for measuring problematic internet use: Implications for pre-employment screening. *Cyberpsychology & Behavior*, 5(4), 331–345. doi:10.1089/109493102760275581

Deloitte. (2015). *Digital media: Rise of on-demand content*. Retrieved April 10, 2018, from www.deloitte.com/in

Digital in 2017: Global Overview. (2017). *We are social*. Retrieved April 14, 2018, from <https://weare-social.com/special-reports/digital-in-2017-global-overview>

Dossier. (2016). Internet usage in Brazil. *The Statistical Portal*. Retrieved August 2, 2018, from <https://www.statista.com/topics/2045/internet-usage-in-brazil/>

Ernst & Young. (2016). *Future of digital content consumption in India*. Retrieved September 13, 2018, from www.ey.com

Farrell, A. D., & Bruce, S. E. (1997). Impact of exposure to community violence on violent behavior and emotional distress among urban adolescents. *Journal of Clinical Child Psychology*, 26(1), 2–14. doi:10.1207/15374424jccp2601_1

Ferguson, D. A. (2012). Online television strategies. In S. T. Eastman & D. A. Ferguson (Eds.), *Media programming* (9th ed., pp. 128-153). Boston, MA: Wadsworth Cengage Learning.

Funk, B. J., Baldacci, B. H., Pasold, T., & Baumgardner, J. (2004). Violence exposure in real-life, video games, television, movies, and the internet: Is there desensitization? *Journal of Adolescence*, 27(1), 23–39. doi:10.1016/j.adolescence.2003.10.005

Gentile, A. D., Lynch, J. P., Linder, R. J., & Walsh, A. D. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of Adolescence*, 27(1), 5–22. doi:10.1016/j.adolescence.2003.10.002

Goldberg, I. (1996). *Internet addiction disorder*. Retrieved 21 June 2018 from <http://www.cog.brown.edu/brochure/people/duchon/humor/internet.addiction.html>

Greenberg, S. B. (1974). Gratifications of Television Viewing and Their Correlates for British Children. In *The Uses of Mass Communication: Current Perspectives on Gratifications Research*. Beverly Hills: Sage.

Greenfield, D. N., & Davis, R. A. (2002). Lost in cyberspace: The web@ work. *Cyberpsychology & Behavior*, 5(4), 347–353. doi:10.1089/109493102760275590

Griffiths, M. (1995). Technological addictions. *Clinical Psychology Forum*, 76, 14-19.

Harrell, B. S. (2000). *Uses and gratification of the Internet* (Unpublished Master Thesis). Texas Tech University.

Harris, M. B., & Williams, R. (1985). Video games and school performance. *Education*, 105(3), 306–309.

Hoffman, L. D., Novak, P. T., & Stein, R. (2012). *The Digital Consumer*. The Routledge Companion to Digital Consumption.

Horsey, D. (2013). ‘Breaking Bad’ bares binge-viewing obsession. *LATimes.com*. Retrieved July 20, 2018, from <http://articles.latimes.com/2013/oct/02/entertainment/la-et-ct-breaking-bad-binge-viewing-20131001>

Huston, A. C., Donnerstein, E., Fairchild, H., Feshbach, N. D., Katz, P. A., Murray, J. P., Rubinstein, E. A., Wilcox, B. L., & Zuckerman, D. M. (1992). *Big world, small screen: The role of television in American society*. University of Nebraska Press.

James, M., & Battaglio, S. (2017). How the big TV networks are adapting to ad-skipping viewers and Google, Snapchat and Facebook. *Los Angeles Times*. Retrieved August 5, 2018, from <http://www.latimes.com/business/hollywood/la-fi-ct-upfront-television-advertising-20170514-story.html#>

Kalus, R. (2016). Time to switch off? When media consumption becomes an addiction. *DW made for minds*. Retrieved July 19, 2018, from <https://www.dw.com/en/time-to-switch-off-when-media-consumption-becomes-an-addiction/a-36738773>

Kandell, J. J. (1998). Internet addiction on campus: The vulnerability of college students. *Cyberpsychology & Behavior*, 1(1), 11–17. doi:10.1089/cpb.1998.1.11

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *Public Opinion Quarterly*, 37(4), 509–523. doi:10.1086/268109

Kim, J., LaRose, R., & Peng, W. (2009). Loneliness as the cause and effect of problematic internet use. The relationship between internet use and psychological well-being. *Cyberpsychology & Behavior*, 12(4), 451–455. doi:10.1089/cpb.2008.0327

KPMG International. (2013). *China, Brazil and Singapore lead consumption of Digital Media and the willingness to pay for it*. Retrieved June 15, 2018, from <https://www.prnewswire.com/news-releases/china-brazil-and-singapore-lead-consumption-of-digital-media-and-the-willingness-to-pay-for-it-187695081.html>

Kubey, R., & Csikszentmihalyi, M. (2002). Television addiction is no mere metaphor. *Scientific American*, 286(2), 74–80. doi:10.1038/scientificamerican0202-74

Lan, C. M., & Lee, Y. H. (2013). The predictors of internet addiction behaviours for Taiwanese elementary school students. *School Psychology International*, 34(6), 648–657. doi:10.1177/0143034313479690

LaRose, R., Lin, C. A., & Eastin, M. S. (2003). Unregulated internet usage: Addiction, habit, or deficient self-regulation. *Media Psychology*, 5(3), 225–253. doi:10.1207/S1532785XMEP0503_01

Lee, M. (2013). *Online streaming sites offer convenience, distraction*. Retrieved June 10, 2018, from <http://niquet.net/life/2013/02/14/online-streaming-sites-offer-convenience-distraction/>

Li, S. N. (2013). *Gratification obtained from television shows on internet TV and conventional TV*. Graduate Theses and Dissertations. Iowa State University.

Lieberman, D. A., Chaffee, S. H., & Roberts, D. F. (1988). Computers, mass media, and schooling: Functional equivalence in uses of new media. *Social Science Computer Review*, 6(2), 224–241. doi:10.1177/089443938800600204

Lotz, A. (2007). *The Television Will Be Revolutionized*. New York University Press.

MarketCast Study Finds TV. (2013, March 8). “Binge-Viewing” Creates a More Engaged Viewer For Fortune Seasons and Not a Bingeing Habit. *PRWeb.com*. Retrieved May 28, 2018, from <http://www.prweb.com/releases/2013/3/prweb10513066.htm>

Marks, I. (1990). Behaviour (non-chemical) addictions. *British Journal of Addiction*, 85(11), 1389–1394. doi:10.1111/j.1360-0443.1990.tb01618.x

Marshall, P. (2009). Screens: Television’s dispersed “broadcast”. In G. Turner & J. Tay (Eds.), *Television studies after TV: Understanding television in the post-broadcast era*. Routledge.

McIlwraith, R. D. (1998). “I am addicted to television”: The personality, imagination and TV watching patterns of self-identified TV addicts. *Journal of Broadcasting & Electronic Media*, 42(3), 371–386. doi:10.1080/08838159809364456

Meerkerk, G. J., Eijnden, R. J. V. D., & Garretsen, H. F. (2006). Predicting compulsive internet use. *Cyberpsychology & Behavior*, 9(1), 95–103. doi:10.1089/cpb.2006.9.95

Meerlo, J. (1954). Television addiction and reactive apathy. *The Journal of Nervous and Mental Disease*, 120(3), 290–291. doi:10.1097/00005053-195409000-00018

Melton, S. (2013). The rise of the second screen: Is the internet overtaking TV? *Brandwatch*. Retrieved April 12, 2018, from <https://www.brandwatch.com/blog/has-the-internet-taken-over-tv/>

Montpetite, J. M. (2014). The internet is changing the definition of television. *The Guardian*. Retrieved June 10, 2018, from <https://www.theguardian.com/media-network/media-network-blog/2014/jun/10/internet-changing-definition-television>

Narang, R., & Trivedi, P. (2016). Challenges and Opportunities of E-Tailing in Emerging Economies. In S. Dixit & A. Sinha (Eds.), *E-Retailing Challenges and Opportunities in the Global Marketplace* (pp. 50–71). IGI Global., doi:10.4018/978-1-4666-9921-2.ch003

Nielsen. (2014). *The Digital Consumer*. Retrieved June, 18, 2018, from <https://www.nielsen.com/content/dam/corporate/us/en/reportsdownloads/2014%20Reports/the-digital-consumer-report-feb-2014.pdf>

Noll, A. M. (2004). Internet television: Definition and prospects. In E. Noam, J. Groebel, & D. Gerbard (Eds.), *Internet Television* (pp. 1–8). Lawrence Erlbaum Associates, Inc.

Palmgreen, P., & Rayburn, J. D. (1979). Uses and Gratifications and Exposure to Public Television: A Discrepancy Method. *Communication Research*, 6(2), 155–179. doi:10.1177/009365027900600203

Pena, L. L. (2015). *Breaking binge: Exploring the effects of binge watching on television viewer reception*. Syracuse University.

Pew Research Center. (2015). *Internet and Technology*. Retrieved July 19, 2018, from <http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015/>

Schechner, S., & Stewart, C. S. (2012). Hulu to create more original shows. *The Wall Street Journal*. Retrieved May 5, 2018, from <http://online.wsj.com/article/SB10001424052970204468004577163162257430538.html>

Shotton, M. A. (1991). The costs and benefits of ‘computer addiction’. *Behaviour & Information Technology*, 10(3), 219–230. doi:10.1080/01449299108924284

Sjoberg, L., & Magneberg, R. (2007). Action sampling of television viewing: A favourite waste of time. *Journal of Media Psychology*, 12. Retrieved April, 25, 2019, from <http://www.calstatela.edu/faculty/sfisco/A%20Favourite%20waste%20of%20Time.html>

Sodano, T. M. (2012). Television’s paradigm (time) shift in time. In *Television narrative: Exploring temporality in twenty-first century programming*. University Press of Mississippi.

State of Digital Downloads. (2017). *Annual survey of Limelight Networks*. Author.

Steinkamp, C. (2010). *Internet television use: motivations and preferences for watching television online among college students* (Thesis). Rochester Institute of Technology.

Stephen, T. A. (2015). The role of digital and social media marketing in consumer behaviour. *Current Opinion in Psychology*, 10(16), 17–21.

Suhail, K., & Bargees, Z. (2006). Effects of excessive internet use on undergraduate students in Pakistan. *Cyberpsychology & Behavior*, 9(3), 297–307. doi:10.1089/cpb.2006.9.297

TiVo Spring Binge Viewing Survey. (2014). *Tivo.com*. Retrieved May 16, 2018 from <https://blog.tivo.com/2014/06/tivo-binge-survey-results-are-in/>

Umstead, R. T. (2014). Binge Nation. *Multichannel News*, 35(16), 12–13.

Wang, J. L., Wang, H. Z., Gaskin, J., & Wang, L. H. (2015). The role of stress and motivation in problematic smartphone use among college students. *Computers in Human Behavior*, 53, 181–188. doi:10.1016/j.chb.2015.07.005

Xue, Y., Dong, Y., Luo, M., Mo, D., Dong, W., Zhang, Z., & Liang, H. (2018). Investigating the impact of mobile SNS addiction on individual's self-rated health. *Internet Research*, 28(2), 278–292. doi:10.1108/IntR-05-2017-0198

Yeomans, M. (2013). Are children consuming too much digital technology. A digital world: what impact does the technology sector have on children? *Guardian Sustainable Business*. Retrieved July 3, 2018, from <https://www.theguardian.com/sustainable-business/children-consuming-too-much-digital-technology>

Young, K. S. (1996). Addictive use of internet: A case that breaks the stereotype. *Psychological Reports*, 79(3), 899–902. doi:10.2466/pr0.1996.79.3.899

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *Cyberpsychology & Behavior*, 1(3), 237–244. doi:10.1089/cpb.1998.1.237

Yuen, C. N., & Lavin, M. J. (2004). Internet dependence in the collegiate population: The role of shyness. *Cyberpsychology & Behavior*, 7(4), 379–383. doi:10.1089/cpb.2004.7.379

Zurawik, D. (2013). The year appointment television died. *BaltimoreSun.com*. Retrieved July 28, 2018 from <http://articles.baltimoresun.com/2013-12-27/entertainment/bal-the-year-appointment-television-died-20131227>

ADDITIONAL READING

Arpaci, I., Kesici, S., & Baloglu, M. (2018). Individualism and internet addiction: The mediating role of psychological needs. *Internet Research*, 28(2), 293–314. doi:10.1108/IntR-11-2016-0353

Bisen, S. S., & Deshpande, Y. M. (2018). Understanding internet addiction: A comprehensive review. *Mental Health Review (Brighton)*, 23(3), 165–184. doi:10.1108/MHRJ-07-2017-0023

Chen, C. S., Wu, M. S., Nguyen, B., & Li, S. (2019). Digitally facilitated newspaper consumption and value co-creation. *The Bottom Line (New York, N.Y.)*, 32(1), 16–46. doi:10.1108/BL-09-2018-0038

Fuentes, C., Hagberg, J., & Kjellberg, H. (2019). Soundtracking: Music listening practices in the digital age. *European Journal of Marketing*, 53(3), 483–503. doi:10.1108/EJM-10-2017-0753

Keegan, S. (2012). Digital technologies are re-shaping our brains: What are the implications for society and the research industry? *Qualitative Market Research*, 15(3), 328–346. doi:10.1108/13522751211232012

Kim, S. (2009). A mobile video service trial: TOEST. *Library Hi Tech*, 27(1), 134–150. doi:10.1108/07378830910942982

Mutum, D. S. (2014). Review Subject: The Routledge Companion to Digital Consumption. *Qualitative Market Research*, 17(4), 495–499.

Watkins, R., & Molesworth, M. (2012). Attachment to Digital Virtual Possessions in Videogames. In R. W. Belk, S. Askegaard, & L. Scott (Eds.), *Research in Consumer Behavior* (pp. 153–170). Emerald Group Publishing Limited. doi:10.1108/S0885-2111(2012)0000014012

KEY TERMS AND DEFINITIONS

Binge Watching: Binge watching means viewing episodes of a show one after the other.

Content: It refers to the storyline, time duration, picturisation, and number of episodes of the shows available on several video streaming portals.

Control: Control refers to an act of ruling something or someone.

Convenience: Convenience means something that makes things easier, comfortable, and more suitable for people.

Digital Consumption: It refers to the sum of entertainment and information media consumed by an individual.

Internet Addiction: It is defined as a compulsive behaviour which makes an individual dependent and disrupts one's professional and personal life.

Internet Television: It refers to all those content which are produced primarily for viewing on the television set via cable or satellite, which is later made accessible online via various applications; content produced primarily for online distribution such as various web-series and user-generated content which can be produced by anybody and made available through YouTube.

Skip Commercials: Skip commercials refer to a feature that makes it possible for viewers to watch shows without advertisements or commercials.

Video Streaming Portals: It refers to a platform or application on which several television shows, series, movies, and videos are available.

Television Use and Consumption of Elderly Americans

8

Robert Andrew Dunn <https://orcid.org/0000-0003-0415-6662>*East Tennessee State University, USA***Stephen W. Marshall** <https://orcid.org/0000-0001-8624-082X>*East Tennessee State University, USA*

INTRODUCTION

While parents, educators, and media critics have done much hand-wringing over the potential effects of modern television on today's youth (Domingues-Montanari, 2017), research suggests that the oldest among us might be more at risk when it comes to troubling screen time; the elderly watch more TV and benefit less from it than younger people (Depp, Schkade, Thompson, & Jeste, 2010). Thus, the purpose of the present research is to offer insight into how the elderly in the United States use, consume, and are affected by television, both in traditional legacy media settings and in newer digital media formats. This entry begins with background information, including definitions and a discussion of the changes in modern television consumption. Next, TV use by senior citizens and the potential effects of that use is discussed. Finally, solutions, recommendations, and potential future research avenues are offered.

BACKGROUND

The focus of this encyclopedia entry is to understand the impact of television consumption on the elderly population. With the advent of digital technologies impacting and transforming how we consumer media it becomes imperative to clearly define the two main concepts of this chapter. For the purposes of this study, the following terms are defined as indicated:

Elderly population: The term “elderly” can have different meanings depending on the source (Weeks, 2013) with the origination of the word coming from the word “elder” and generally means old (Online Etymology Dictionary, n.d.). The United States Social Security office considers people over the age of 66 eligible for full retirement with full benefits (Borland, 2017). However, most of the research presented here focused on this population seems to reference people over the age of 60 (e.g. Fouts, 1989 or Eggermont & Vandebosch, 2001) as being part of the “aging” population. This population is of particular importance because it is an under-researched growing population. The U.S. Census Bureau as cited in the Population Reference Bureau (2015) projects the number of Americans aged 65 and older is projected to double to 95 million by 2060 - sharing nearly a quarter of the population (Mather, Scommegna, & Kilduff, 2015).

Television Consumption: With nearly 80% of Americans watching some form of television on a daily basis and it is considered “America’s favorite pastime.” For the purpose of this study, the American

DOI: 10.4018/978-1-7998-3479-3.ch074

Time Use Survey (ATUS) from the United States Bureau of Labor Statistics (BLS) defines “watching TV” as the definition of television consumption, “As defined in the ATUS, “watching TV” refers to any time people said their main activity involved watching TV, videos, or movies. This includes the time they spent watching live programming, viewing DVDs, and streaming shows on their TV sets, computers, and portable devices. It does not include time spent viewing movies at a theater” (Krantz-Kent, 2018).

The Changing Landscape of Traditional Television Viewing

Television consumption, as with most media consumption, has changed drastically in the last 20 years (Einav & Wolf, 2019). Internet use has shown to cut into television time, particularly for the most experienced internet users (Cole, 2003). Younger audiences, those born in 1980 and after, are driving these changes, as they expect more options as well as a more personalized media consumption experience (Einav & Wolf, 2019). Almost half of all adults, 22-45 years old, are not watching traditional TV (Poggi, 2017). What has developed is a new age of “cord cutters,” those who end their pay TV subscription from cable and satellite, and “cord nevers,” those who never paid for those services to begin with and never intend to. These two groups are also responsible for the rise in over-the-top (OTT) television services, TV delivered via the internet through online services such as Netflix and Hulu (Einav & Wolf, 2019, para. 19-20). By 2017, according to Parks Associates, 69 percent of U.S. broadband households were OTT subscribers, with more than half paying for more than one service. This trend is expected to continue as about 75% of all TV households in the U.S. have a connected device, with the average household having 3.9 connected devices (Frankel, 2019).

Content providers for streaming services have exploded in recent years as well. Netflix is probably best known for originally delivering movies to homes via mail to today becoming a powerful streaming service. In fact, some even reference “the Netflix Effect” as being when a new series catapults the fame of a program or actor overnight (Morgan, 2019). The sharable nature of the availability of this content creates accelerated viewing and popularity like nothing seen in the past. In addition to Netflix and Hulu, the market is flooded with previous cable channel content providers (TBS, TNT, CNN) and premium content providers such as HBO – each creating streaming applications and online portals. For example, a recent acquisition of Time Warner by AT&T has created a new company called WarnerMedia, aimed to compete with the likes of Netflix (Sims, 2019). Just as the devices have changed so has the production and distribution of content. These changes will undoubtedly continue for the foreseeable future as digital technology continues to evolve.

TELEVISION CONSUMPTION TRENDS OF ELDERLY AMERICANS

Despite the hype of digital, networked, and mobile technology, television is still the most voracious consumer of time for older Americans. Research suggests that as people age, their consumption of television increases. This is particularly true among senior citizens, who rely on TV more than any other form of media (Fouts, 1989; Nielsen, 2014). In fact, the average American over 65 spent 30 more minutes per day watching television when compared to the year 2003 (Kopf, 2019). According to the Bureau of Labor Statistics (2018), of all age groups, older Americans spend the most time watching television. Trending from 2013-2017, the average hours per day watching television for 55-64 years of age was 2.9 hours per day. Americans over the age of 65 spent an hour more watching an average of 4.1 hours per day while those over the age of 75 watch an average of 4.5 hours per day. This trended significantly

higher one year later with those 65+ watching an average of 6 hours and 51 minutes per day as measured in the third quarter of 2018 (Marketing Charts, 2019). Comparatively, Americans average just under two hours per day across all age breaks under 45 years old. The power of television is profound with Americans over the age of 65 spending more time watching TV than reading, socializing, relaxing and all other leisure activities combined. There is no doubt senior citizens and those approaching senior citizen status account for a large percent of the television viewing audience. The most popular television show for the last five years, as of 2014, has been CBS's NCIS, a military-based police procedural, which has a median viewer age of 60 (Kang, 2014).

In today's digital media environment, time-shifting television has become the norm with cable, satellite, and streaming services allowing viewers to watch "on demand" and "binge" on consecutive programs in a series. Streaming (time-shifting) is on the increase with all television viewers but elderly viewers are beginning to participate in the trend. In 2017, a study by Wall Street investment bank Barclays found viewing on OTT sources such as an internet-connected smart TV or device has increased 45% among 55-64 year olds and increased by 36% among those over the age of 65 (Kiger, 2017). However, as of 2017, almost half of people 65 and up still do not have broadband. And the tech gap is exaggerated for those who are entering older age at 65 and those who are 75 and older. For broadband, 66 percent of people 65 to 69 years old are subscribers. That compares to only 41 percent for ages 75 to 79 and a mere 28 percent for ages 80 and older. For smartphones, 59 percent of people 65 to 69 had a smartphone. That compares to only 31 percent for ages 75 to 79 and a paltry 17 percent for people 80 and older (Anderson & Perrin, 2017). OTT and digital media will continue to elude the majority of the oldest Americans until those adoption rates are improved. And while the internet does have the potential to cut into television time (Cole, 2003), the oldest are the least likely to be connected, with only 44 percent of people 80 and up using the internet (Anderson & Perrin, 2017).

Television and News

Television news has historically been a powerhouse for informing the public and for decades, prior to internet, lead in "breaking" news and coverage of live events. However, the internet and digital media has vastly changed how the American public is informed – with the exception of elderly Americans. The Pew Research Center (2016) reported most people age 65 and older (85%) turn to television for their news with only 48% stating they get their news from newspapers. When comparing television news by age, the data shows a stark difference in age with 27% of 18-29 year olds and 45% of 30-49 year olds "often" getting their news from television. The digital divide among generations can be illustrated with nearly 50% of 18-49 year olds "often" getting their news online compared to only 20% Americans 65 and over. Thus, television serves as a vital link between seniors and information about the world around them.

When comparing news consumption across news outlets such as local, national network (i.e. CBS, NBC, etc.), and cable, elderly Americans are still consuming the most as well. Only 18% of 18-29 year olds get their news from local television compared to 57% of those aged 65+. In a similar fashion, elderly Americans are "often" getting their news from network news (49%) and cable news (58%) compared to 18-29 year olds often getting their news with 8% of network news and 10% of cable news, respectfully.

Television and Quality Perceptions

Given the amount of time seniors spend with TV, one would think they find great enjoyment from the pastime. However, research by Depp, Schkade, Thompson, and Jeste (2010) contradicts such a deduc-

tion. According to their study, adults 65 and up do indeed watch more TV, but enjoy it significantly less than their younger counterparts. Older adults did not receive the stress-reduction benefits of television consumption compared to their younger counterparts. Moreover, television consumption among older adults coincided with greater sadness. The authors speculated that seniors might watch TV more because they have more free time and because it is an easy and accessible activity when compared to others, such as exercise. They also suggested that TV consumption by older adults could be related to habitual/addictive cycles, use despite a diminishment of returns. In a similar vein, researchers found that older people who watch more television reported greater dissatisfaction with their own lives. This finding was attributed to a possible disconnect between the viewer's financial status versus the wealth, riches, and material possessions the viewer sees on TV (Rahtz, Sirgy, & Meadow, 1988).

Given the great amount of time that elderly adults spend with television, there is an understandable concern about the effects that medium has on them. Vickers (2007) argued that, historically, the portrayal of older Americans on TV has not been positive. She suggested that the skew toward younger talent and storylines that appeal to younger audiences effectively tells older audiences that they are not valued in the media landscape. She also noted that the recent focus on anti-aging cosmetics and surgery is also troublesome. She pointed to research that showed that the more seniors watch TV, the less positive they are about growing old.

EFFECTS OF TV ON THE ELDERLY

Health Impact of Television on the Elderly

Activity is understood to be a part of a healthy lifestyle, but for the elderly, being active can become more and more challenging with age. Inoue et al. (2012) found that people who watched less television were also less likely to be overweight or obese, a finding that coincides with television viewing as contributing to a sedentary lifestyle.

In particular, [sedentary behavior (SB)], or proxy measures such as television viewing, have been shown to be associated with an increased risk of obesity, cancer, metabolic syndrome, type 2 diabetes, and all-cause mortality in various populations. There has been very limited research focusing on adults aged ≥65 years despite evidence to suggest older adults represent the most sedentary age group, spending approximately 60% to 70% of their waking hours in SB. (Gennuso, Gangnon, Matthews, Thraen-Borowski, & Colbert, 2013, para. 5)

In fact, that is exactly what research from de Rezende, Rey-López, Matsudo, & do Carmo Luiz (2014) showed. The researchers found that the elderly spend the majority of their day participating in sedentary activities, such as watching television. And that sedentary lifestyle, the researchers found, is related to a greater risk in mortality in older adults.

Physical Limitations

Watching television is traditionally a sedentary activity and with the elderly, mobility can become problematic. A study by DiPietro, Jin, Talegawkar and Matthews (2018) looked at the association between sedentary time and physical activity with mobility. Analyzing data from 134,269 people in the National

Institutes of Health (NIH)-American Association of Retired Persons (NIH-AARP) Diet and Health Study, the researchers found a reduction in sedentary time combined with an increase in physical activity can lead to or at least maintain mobility function in older age. The study found, “Within all levels of physical activity, increasing amounts of TV time increased the odds of mobility disability in a dose–response manner, with almost a curvilinear accelerated odds of disability observed with increasing TV time in combination with the lowest levels of physical activity” (para. 14). A study done outside of the United States found older adults who watched less television were less likely to be overweight or obese independent of their participation in moderate-to-vigorous physical activity (Inoue et al., 2012). Reed et al. (2017) conducted a 12-year television viewing study examining performance-based measures of physical function concluding patterns of sedentary behavior can be a determinant of muscle strength in later life.

Dementia

According to an international study by Fancourt and Steptoe (2019), watching too much television can increase the likelihood of dementia.

Dementia is an overall term for diseases and conditions characterized by a decline in memory, language, problem-solving and other thinking skills that affect a person’s ability to perform everyday activities. Memory loss is an example. Alzheimer’s is the most common cause of dementia. (Alzheimer’s Association, 2019, para. 1)

Using data from the *English Longitudinal Study of Aging*, longitudinal associations between television watching and cognition six years later were compared while controlling for other factors such as demographics, depression, health behaviors, socio-economic, etc., suggesting more than 3.5 hours per day of television viewing is related to cognitive decline (McDermott, 2019). According to researcher Daisy Fancourt, there has been a dearth of studies dedicated to the effects of TV on the minds of the oldest viewers even though scientists have speculated a connection between too much screen time and dementia for decades. Older Americans, therefore, are encouraged to incorporate more activities like reading books, listening to music, and socializing with friends to balance potential ill effects of TV (McDermott, 2019).

Psychological Impact of Television

Fouts (1989) explicated a number of potential motivations for why the elderly are so attached to television as a medium. Those motivations included “information-seeking,” “entertainment,” “topics for conversation,” “to help schedule their daily activities,” “for substitute companionship,” “sharing the viewing experience,” correlating with one’s “self-concept,” and correlating with one’s “sensory decline” (Fouts, 1989, pp. 570-571). Fouts argued that the more motivations the elderly had for watching television, the more they would use it. While all of these motivations might not be unique to the elderly, there does seem to be a nexus of issues driving the elderly to the screen, as getting older means losing loved ones and not being able to get around easily.

The loneliness motivation is something that continues to be found in more recent research. Eggermont and Vandebosch (2001) studied 284 people 60 years and older. First, their consumption was nearly double the general population. “Parasocial interaction, loneliness relief, passing time, boredom avoidance, and activity substitution are obvious viewing motivations” (p. 10). The life circumstances of

the elderly, such as a lack of social connectivity, seemed to drive television use and a connection with the characters they saw on the screen.

Nguyen, Wittink, Murray, and Barg (2008) cited previous research potentially connecting television use with depression in the elderly. The researchers, however, set out to study what people 65 and older thought themselves about the relationship between television and depression. Elderly respondents said that TV use could cause them to realize they were depressed because they were over using it, that TV could be a coping mechanism for dealing with depression, or that TV could be a trigger for depression or could make depression worse.

What the elderly think about TV is also what drove research for Van Der Goot, Beentjes, and Van Selm (2012). In studying Dutch people 65 and up, the researchers found that television viewing had changed for respondents based on available free time, based on social and familial changes, and based on TV content itself. Some people watched more TV because retirement offered that ability, though others watched less because they did not need TV for relaxation from their job. Some watched more TV after a spouse's death. Some watched less TV to avoid painful memories of those who they have lost. Finally, some watched new TV genres because they no longer had a spouse to compromise with on choice or because they wanted to stay mentally sharp. Others found they had to turn off some genres as they had become more sensitive to the content. While it is true that people might turn to television as they age, such research shows this group is far from a monolith in terms of both motivation and behavior.

SOLUTIONS AND RECOMMENDATIONS

As the oldest Americans are both the most sedentary and least connected (Anderson & Perrin, 2017; de Rezende, Rey-López, Matsudo, & do Carmo Luiz, 2014), more outreach work must be done to ensure such people are not falling through the cracks both in terms of digital connectivity and overall health. Social networks do seem to have the capacity to be beneficial to the elderly, leaving those who use it feeling overall healthier, suffering from fewer chronic health issues, and experiencing less depression and loneliness (Frank, 2018).

The elderly can often find free programs that offer classes to train them in computers. Classes can also be found at senior citizen centers, libraries, and colleges (Frank, 2018). These courses offer real-world solutions for not just getting the elderly online, but for turning them onto the newest technologies and media trends.

Passively offering courses on computers might be enough to push the elderly to Google their medication, try out social media, or build a website (Frank, 2018), but it might not be enough to change literally decades of television habituation. Where are the classes for OTT services such as Netflix and Hulu? How many courses offer seniors a chance to go hands-on with streaming devices, such as a Roku Player or an Amazon Fire TV Stick? Classes do exist to teach seniors on how to use smartphones and tablets (Orlov, 2016), but such offerings are for older people who are already motivated to learn these new technologies. Who is evangelizing these technologies to the elderly? Who is showing seniors how much more convenient streaming can be than traditional television? How liberating watching the news on a tablet anywhere can be? How one can watch the game on TV and connect with like-minded sports fans with a smartphone to enjoy a second-screen experience (Poltrack, 2012)? While having classes on computers for the elderly is clearly a good thing, there also needs to be classes that align with how people actually spend their leisure time today. And waiting for students to show up is not the solution here. Again, outreach is the key.

FUTURE RESEARCH DIRECTIONS

Seniors lag behind younger Americans when it comes to overall technology adoption. However, seniors are adopting new digital technology (mobile, tablets, etc.) at an increasing rate. As consistent with younger audiences, adoption of new technology is sure to impact traditional television viewing as well as the way television content is consumed. According to the Pew Research Center (2014) more than 59% of seniors reported they go online (six percent increase from the previous year) with 47% having access to high-speed broadband connections. New technology adoption increases even more when considering a subgroup of seniors who have a higher social status, education or financial liquidity. As stated in the Pew Research Center (2014) study, “Younger, higher-income, and more highly educated seniors use the internet and broadband at rates approaching-or even exceeding-the general population.” This is consistent with Rogers’ (2003) diffusion of innovations perspective on early adopters as a subgroup of the senior population. As the diffusion curve moves, more adoption and changes in television consumption should be anticipated. Research opportunities are ripe to review how and at what rate older Americans adopt technology as well as how their television consumption, based on this adoption, has changed.

Additionally, television program content providers are moving access to broadband, on-demand, and DVR-based viewing platforms in homes adopting newer technology such as smart televisions. Approximately 32% of all televisions in U.S. households are now considered “smart,” allowing for streaming of content via applications from the internet, with 14% of U.S. adults watching daily. Combine this with nearly half of all U.S. homes having at least one stand-alone streaming device (Frankel, 2019) and it is clearly understood how content delivery is changing. Streaming is becoming mainstream, and thus understanding how the elderly are participating in this shift, or not, is of great importance.

Online content providers such as Hulu, Netflix and Amazon are partnering with traditional broadcast and cable networks to provide programming in new ways as well. Traditional networks are also creating interactive experiences in via web and mobile viewing platforms. Increased participation in social media from audiences, television content brands and seniors is sure to impact viewing habits. Also, as generations’ age, their familiarity with technology will change. Additionally, the impact of technology on television consumption is sure to change the traditional viewing model for seniors in the future. There is no doubt media delivery (aka binge watching) will continue to change consumption habits and there is currently no research examining how this behavior will impact the physical or emotional wellness of the elderly.

CONCLUSION

This entry communicates the television consumption habits of seniors currently differ from the general population. However, technology will continue to have an increased impact in viewing habits in the future. Television consumption in the senior population is a ripe area for research since people are living longer and availability of television via digital technology is becoming more mainstream. While there is significant information in terms of industry trend data there seems to be a general deficit in scholarly data examining the impact of television consumption as well as technology delivery on the senior generation.

REFERENCES

- Alzheimer's Association. (2019). *What is Dementia?* Retrieved from Alzheimer's Association website: <https://www.alz.org/alzheimers-dementia/what-is-dementia>
- Anderson, M., & Perrin, A. (2017, May 17). Tech Adoption Climbs Among Older Americans. *Pew Research Center*. Retrieved July 30, 2019, from Pew Center Research website: <https://www.pewinternet.org/2017/05/17/tech-adoption-climbs-among-older-adults/>
- Borland, J. (2017). 2017 Brings New Changes to Full Retirement Age. *Social Security Matters*. Retrieved from <https://blog.ssa.gov/2017-brings-new-changes-to-full-retirement-age/#more-2247>
- Cole, J. I. (2003). *The UCLA Internet report: Surveying the digital future, year three*. UCLA Center for Communication Policy. Retrieved from https://www.digitalcenter.org/wp-content/uploads/2013/02/2003_digital_future_report-year3.pdf
- de Rezende, L. F. M., Rey-López, J. P., Matsudo, V. K. R., & do Carmo Luiz, O. (2014). Sedentary behavior and health outcomes among older adults: A systematic review. *BMC Public Health*, 14(1), 333. doi:10.1186/1471-2458-14-333 PMID:24712381
- Depp, C. A., Schkade, D. A., Thompson, W. K., & Jeste, D. V. (2010). Age, Affective Experience, and Television Use. *American Journal of Preventive Medicine*, 39(2), 173–178. doi:10.1016/j.amepre.2010.03.020 PMID:20621265
- DiPietro, L., Jin, Y., Talegawkar, S., & Matthews, C. E. (2018). The Joint Associations of Sedentary Time and Physical Activity With Mobility Disability in Older People: The NIH-AARP Diet and Health Study. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 73(4), 532–538. doi:10.1093/gerona/glx122 PMID:28958064
- Domingues-Montanari, S. (2017). Clinical and psychological effects of excessive screen time on children. *Journal of Paediatrics and Child Health*, 53(4), 333–338. doi:10.1111/jpc.13462 PMID:28168778
- Eggermont, S., & Vandebosch, H. (2001). Television as a substitute: Loneliness, need intensity, mobility, life-satisfaction and the elderly television viewer. *Communicatio*, 27(2), 10–18. doi:10.1080/02500160108537902
- Einav, G., & Wolf, H. (2019). *From Interactive Television to Digital Health: Insights for Industries in Transformation*. doi:10.1108/978-1-78973-619-920191002
- Fancourt, D., & Steptoe, A. (2019). Television Viewing and Cognitive Decline in Older Age: Findings from the English Longitudinal Study of Ageing. *Scientific Reports*, 9.
- Fouts, G. T. (1989). Television use by the elderly. *Canadian Psychology*, 30(3), 568–577. doi:10.1037/h0079824
- Frank, D. (2018, July 24). *Tech Training Helps Older Americans Socialize*. Retrieved August 1, 2019, from AARP website: <https://www.aarp.org/home-family/personal-technology/info-2018/technology-training-for-older-adults.html>
- Frankel, D. (2019). Smart TV Penetration in U.S. Now up to 32%. *Multichannel News*. Retrieved from <https://www.multichannel.com/news/32-percent-of-us-tvs-are-smart-tvs>

- Gennuso, K. P., Gangnon, R. E., Matthews, C. E., Thraen-Borowski, K. M., & Colbert, L. H. (2013). Sedentary behavior, physical activity, and markers of health in older adults. *Medicine and Science in Sports and Exercise*, 45(8), 1493. Retrieved from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5764165/>
- Inoue, S., Sugiyama, T., Takamiya, T., Oka, K., Owen, N., & Shimomitsu, T. (2012). Television Viewing Time is Associated with Overweight/Obesity Among Older Adults, Independent of Meeting Physical Activity and Health Guidelines. *Journal of Epidemiology*, 22(1), 50–56. doi:10.2188/jea.JE20110054 PMID:22156288
- Kang, C. (2014, September 24). The power of traditional TV: “NCIS” and its older audience deliver gold for CBS. *The Washington Post*. Retrieved from https://www.washingtonpost.com/business/technology/the-power-of-traditional-tv-ncis-and-its-older-audience-deliver-gold-for-cbs/2014/09/24/c4e0f92a-3f4d-11e4-b03f-de718edeb92f_story.html
- Kiger, P. (2017, July 21). *Are you Ready to Jump Into TV Streaming?* Retrieved from AARP website: <https://www.aarp.org/entertainment/television/info-2017/older-adults-watching-tv-online-more-fd.html>
- Kopf, D. (2019). Old People in the US are Watching a lot more TV. *Quartz*. Retrieved from <https://qz.com/1563911/who-watches-the-most-tv/>
- Krantz-Kent, R. (2018). Television, Capturing America’s Attention at Prime Time and Beyond. *American Time Use Survey (ATUS)*. Retrieved from <https://www.bls.gov/opub/btn/volume-7/television-capturing-americas-attention.htm>
- Marketing Charts. (2019, April 3). *The State of Traditional TV: Updated with Q3 2018 Data*. Retrieved from <https://www.marketingcharts.com/featured-105414>
- Mather, M., Scommegna, P., & Kilduff, L. (2015). Fact Sheet: Aging in the United States. *Population Reference Bureau*. Retrieved from <https://www.prb.org/aging-unitedstates-fact-sheet/>
- McDermott, N. (2019, February 28). Watching 3.5 Hours of TV a Day Increases Risk of Dementia: Study. *New York Post (The Sun)*. Retrieved from <https://nypost.com/2019/02/28/watching-3-5-hours-of-tv-a-day-increases-risk-of-dementia-study/>
- Morgan, B. (2019, February 19). What is the Netflix Effect? *Forbes*. Retrieved from <https://www.forbes.com/sites/blakemorgan/2019/02/19/what-is-the-netflix-effect/#267b26435640>
- Nguyen, G. T., Wittink, M. N., Murray, G. F., & Barg, F. K. (2008). More Than Just a Communication Medium: What Older Adults Say About Television and Depression. *The Gerontologist*, 48(3), 300–310. doi:10.1093/geront/48.3.300 PMID:18591355
- Nielsen. (2014). *Shifts in Viewing: The Cross-Platform Report*. Author.
- Online Etymology Dictionary. (n.d.). Retrieved from <https://www.etymonline.com/word/elderly>
- Orlov, L. (2016, March 8). *Finding training for seniors with new smartphones, tablets and more*. Retrieved August 1, 2019, from Aging in Place Technology Watch website: <https://www.ageinplacetechnology.com/blog/finding-training-seniors-new-smartphones-tablets-and-more>
- Parks Associates. (2017, November 30). *More than one-half of OTT-subscribing households pay for multiple OTT video services*. Retrieved July 28, 2019, from <http://www.parksassociates.com/blog/article/pr-11302017>

Pew Research Center. (2014). *Older Adults and Technology Use*. Retrieved from <https://www.pewinternet.org/2014/04/03/older-adults-and-technology-use/>

Pew Research Center. (2016). *The Modern News Consumer*. Retrieved from https://www.journalism.org/2016/07/07/the-modern-news-consumer/pj_2016-07-07_modern-news-consumer_1-01/

Poggi, J. (2017, September 22). Nearly Half of Millennials and Gen Xers Don't Watch Any Traditional TV: Study. *AdAge*. Retrieved July 28, 2019, from <https://adage.com/article/media/half-young-consumers-watching-content-traditional-tv-study/310564>

Poltrack, A. (2012, December 22). *The second screen explained: What it is and how it works*. Retrieved August 1, 2019, from Digital Trends website: <https://www.digitaltrends.com/home-theater/second-screen-explained/>

Rahtz, D. R., Sirgy, M. J., & Meadow, H. L. (1988). Elderly Life Satisfaction and Television Viewership: An Exploratory Study. In J. Micheal (Ed.), *NA - Advances in Consumer Research*, 15 (pp. 141–145). Association for Consumer Research.

Reid, D., Healy, G. N., Daly, R. M., Baker, P., Eakin, E. G., Dunstan, D. W., Owen, N., & Gardiner, P. A. (2017). Twelve-Year Television Viewing Time Trajectories and Physical Function in Older Adults. *Medicine and Science in Sports and Exercise*, 49(7), 1359–1365. doi:10.1249/MSS.0000000000001243 PMID:28240701

Rogers, E. (2003). *Diffusion of Innovations* (5th ed.). Simon & Schuster Inc.

Sims, D. (2019). The Wild-West Era of Streaming TV is Ending. *The Atlantic*. Retrieved from <https://www.theatlantic.com/entertainment/archive/2019/06/warnermedia-and-new-streaming-tv-era/591373/>

Van der Goot, M., Beentjes, J. W. J., & Van Selm, M. (2012). Meanings of television in older adults' lives: an analysis of change and continuity in television viewing. *Ageing and Society*, 32(1), 147–168.

Vickers, K. (2007). Aging and the media: Yesterday, today, and tomorrow. *Californian Journal of Health Promotion*, 5(3), 100–105. doi:10.32398/cjhp.v5i3.1256

Weeks, L. (2013, March 14). An Age-Old Problem: Who is 'Elderly'? *National Public Radio*. Retrieved July 14, 2019 from <https://www.npr.org/2013/03/12/174124992/an-age-old-problem-who-is-elderly>

ADDITIONAL READING

Depp, C. A., Schkade, D. A., Thompson, W. K., & Jeste, D. V. (2010). Age, Affective Experience, and Television Use. *American Journal of Preventive Medicine*, 39(2), 173–178. doi:10.1016/j.amepre.2010.03.020 PMID:20621265

Gross, J. J., Carstensen, L. L., & Pasupathi, M. (1997). Emotion and Aging: Experience,

Gross, J. J., Carstensen, L. L., Pasupathi, M., Tsai, J., Götestam Skorpen, C., & Hsu, A. Y. C. (1997). Expression, and Control. *Psychology and Aging*, 12(4), 590–599. doi:10.1037/0882-7974.12.4.590 PMID:9416628

Harwood, J. (1999). Age identification, social identity gratifications, and television viewing. *Journal of Broadcasting & Electronic Media*, 43(1), 123–136. doi:10.1080/08838159909364479

- Harwood, J., & Anderson, K. (2002). The presence and portrayal of social groups on prime-time television. *Communication Reports*, 15(2), 81–97. doi:10.1080/08934210209367756
- Mroczek, D. K., & Kolarz, C. M. (1998). The Effect of Age on Positive and Negative Affect: A Developmental Perspective on Happiness. *Journal of Personality and Social Psychology*, 75(5), 1333–1349. doi:10.1037/0022-3514.75.5.1333 PMID:9866191
- Peel, H. P., McClure, R. J., & Bartlett, H. P. (2005). Behavioral Determinants of Healthy Aging. *American Journal of Preventive Medicine*, 28(3), 298–304. doi:10.1016/j.amepre.2004.12.002 PMID:15766620
- Reid, K. (1989). Studies Analyze Elderly Use of Television. *Media and Values*, 45.
- Schreiber, E., & Boyd, D. (1980). How the Elderly Perceive Television Commercials. *Journal of Communication*, 30(1), 61–70. doi:10.1111/j.1460-2466.1980.tb01770.x PMID:7372845
- Williams, D., Raynor, H., & Ciccolo, J. T. (2008). A Review of TV Viewing and its Association with Health Outcomes in Adults. *American Journal of Lifestyle Medicine*, 2(3), 250–259. doi:10.1177/1559827608314104
- Wonneberger, A., Schoenbach, K., & Meurs, L. (2009). Dynamics of Individual Television Viewing Behavior: Models, Empirical Evidence, and a Research Program. *Communication Studies*, 60(3), 235–252. doi:10.1080/10510970902955992

KEY TERMS AND DEFINITIONS

Cord-Cutting: The act of canceling a subscription to cable or satellite to consume television content via online content providers such as Netflix, Hulu, or others.

Dementia: A general term for a decline in mental ability severe enough to interfere with daily life. Memory loss is an example. Alzheimer's is the most common type of dementia.

DVR Technology: Digital video recorder is an electronic device that is used to record and play back video content. Many cable and satellite services provide DVR technology as part of the consumer package.

Elderly: People over the age of 60 years old.

Online Content Providers: Text, video, and audio delivered via the internet.

OTT: Over the top television is any television content delivered via an internet connection. Examples of OTT providers are Netflix, Hulu, Amazon Video, etc.

Penetration: Percentage or share of an adopted technology.

Second-Screen Experience: The use of a screened device, such as a smartphone or tablet, while watching another screen, such as a TV or a computer, to enrich the viewing of a program.

Sedentary Behavior (SB): Behaving in a manner that is not physically active, such as sitting and watching television and not standing or moving one's body.

Streaming: The digitization of content delivered via internet. Often also referenced as streaming TV, online TV, or internet TV. Examples of this include Netflix, Hulu, Amazon Video, etc.

Television Consumption: Any time spent watching live or recorded television, viewing DVDs, and streaming shows on their TV sets, computers, and portable devices.

Time-Shifting: Recording a program to a storage medium to be consumed at a later date.

Social Media and Social Movements: Strengths, Challenges, and Implications for the Future

Sheldondra J. Brown

Southeast Missouri State University, USA

Grace M. Babcock

North Carolina State University, USA

Monica Bixby Radu

Southeast Missouri State University, USA

INTRODUCTION

Collective action to advance social change is part of cultural tradition in the United States. Throughout American history, many Americans have advocated for both structural and cultural changes through charismatic leadership and collective action. Regardless if the goal has been for women's right to vote, free lunches for all students, or access to clean water for all Americans, having the freedom to protest and assemble are important components for any democratic society (Carroll & Hackett, 2006). Additionally, the creation of social movements has an underlying goal to incorporate new actors in the political process (Garrett, 2007).

Because of the increasing importance of the link between social media and social movements, recent research attempts to bridge the gap between media studies and sociological research on social movements. For example, Gerbaudo (2012) explores the culture of social movements in his book "Tweets and the Streets: Social Media and Contemporary Activism" arguing that the media provides a channel through which social movements are organized and further developed. He also argues that social media allows individuals to develop connections to others who may have similar interests and causes. However, Morozov (2011) points out that social media facilitates "slacktivism", that is activism for "slackers." Morozov argues that individuals may perceive they are fighting injustice by joining a Facebook group or retweeting, but their actions may have little to no political or social impact.

Therefore, questions remain unanswered. Does social media help facilitate activism that leads to social change? What are the strengths and limitations of social media in creating and maintaining a social movement? We explore these questions and others in our article paying particular attention to the #Blackslivesmatter (BLM) social movement, which according to Lebron (2017) is a powerful force demanding social change in America. The hashtag #Blacklivesmatter was created in July 2013 in response to George Zimmerman's acquittal for the second-degree murder of an African American teenager, Trayvon Martin (Freelon, McIlwain, & Clark, 2016). The watershed moment for this social movement was the fatal police shooting of African American youth, Michael Brown who was shot by a white police officer in Ferguson, Missouri on August 9, 2014 (Anderson & Hitlin, 2016). The authors of this article focus on this social movement because both scholars and activists suggest that BLM is responsible for increasing social awareness regarding police brutality towards Black Americans (Garrett, 2007).

DOI: 10.4018/978-1-7998-3479-3.ch075

BACKGROUND

On February 1, 1960, four freshmen students from North Carolina Agricultural and Technical College went into a store in Greensboro, NC and sat down at a lunch counter that had been deemed for whites only. The four students were not served but remained seated at the whites-only counter until the store closed. Although the student sit-ins appeared to be an unplanned event, there was a great amount of planning that took place before the sit-ins occurred. All four of the students were members of an NAACP Youth Council and had been in contact with other activists who had participated in sit-ins in the late 1950s (Polletta, 2006). The actions of the four students—Ezell Blair Jr., David Richmond, Franklin McCain, and Joseph McNeil—suggested a spontaneous, impulsive form of resistance, rather than a planned social movement. Consequently, Polletta (2006) argued that the spontaneous nature of the sit-ins suggested a powerful moment when a group of individuals suddenly become a *collective force* undertaking social injustice. Yet, these events were part of a much larger, powerful, and historically significant form of collection action in the United States: The Civil Rights Movement.

What is a Social Movement?

Social movements are the collective action of individuals who come together for the purpose of achieving a goal that challenges society's cultural or structural beliefs. Goodwill and Jasper (2009) describe social movements as “being conscious, concerted, and sustained efforts by ordinary people to change some aspect of their society by using extra-institutional means” (p. 5). Formal organizations are created because of social movement while others rely on the informal networks. Therefore, using a rigorous definition for social movements may exclude the work of less mainstream causes.

Scholars identify four stages of social movements, including emergence, coalescence, bureaucratization, and decline (Jones-Eversley et al., 2017). Jones-Eversley and colleagues (2017) argued that these four stages can be applied to both the Civil Rights Movement and BLM. The first stage, *emergence* refers to when the social movement has little organization for how to redress their dislike for a certain policy or social condition. During this stage, people are acting as individual agents not as a collective unit. For example, a person may send a letter or email a representative to voice his or her opinion or they may post their view on a popular social media site. Once in the second stage, *coalescence*, like-minded individuals gather together to focus their message. The movement is more organized to state grievances as a group towards a person, policy, or social standard. In the online context, this stage may have a hashtag (i.e., #metoo, #occuywallstreet) that gains the public's attention. The next stage, *bureaucratization*—the foundation of the social movement—is organized with political power. Now the social movement has moved away from demonstrations of awareness to having a formal organization working towards their collective goals. The final stage is *decline*. Decline occurs due to success, failure, or a disinterest in the movement. Applying these stages to one of the most significant social movements in American history, the Civil Rights Movement progressed through these stages from Rosa Parks refusal to give up her seat, the boycott of the Montgomery bus station, NAACP legal involvement, and finally the case Browder v. Gayle to end racial segregation on Alabama buses (Browder v Gayle).

What is Social Media?

Social media is an umbrella term used to explain the networking of communication and collaboration online (Friedland, 2013). The general formats rely on individuals connecting and networking with others

in multiple forms and functions. Several networking sites have platforms that serve different purposes of sharing information. Worldwide, the social media website with the most users is Facebook. Globally, the United States has the second-greatest number of Facebook users (190 million), following behind India (270 million) (Statista, 2019). Facebook is best known for personal interactions to stay connected with friends and family. In the United States, as of July 2019, there were over 48 million Twitter users (Statista, 2019). Twitter users reach a large audience of “followers” for fast, short information. Instagram relies on interactions between users and the follower sharing pictures, videos, and visual media. LinkedIn provides users the ability to maintain formal networks for employment and professional relationship. In addition to this being informal networks online, there are interactions with state officials, celebrities, businesses, and powerful individuals.

Using social media to communicate is breaking down barriers of reaching potential followers and influential individuals. Social media provides an open access to dialogue that is decentralized (Freberg, 2013). It does not rely on one-way dissemination of content and the cost to create content bare little or any expense. Unlike traditional platforms of communication (e.g., newspapers), social media has little delay, which allows it to break barriers of time and location. For many social media users, checking social media websites is part of their daily routine, as Pew Research (2018) estimates that three-quarters of Facebook users and six to ten Instagram users visit these sites at least once a day.

There are various types and purposes of social media. As shown in Table 1, social networking websites, such as Myspace (founded in 2003), Facebook (founded in 2004), and Twitter (founded in 2006) are typically used as a space for social interaction with family, friends, and other followers. A “follower” on social media refers to someone who subscribes to receive a person’s updates (Beal, 2019). Social networking websites can be used to both maintain and develop new social connections. Social review websites, such as Yelp and TripAdvisor are utilized to share reviews of businesses with potential customers. On image sharing websites, such as Instagram and Pinterest (both founded in 2010), users can share images with their friends and followers. Video hosting websites, including Vimeo and YouTube

Table 1. Summary of Types and Purpose of Various Social Media Platforms (2000-Present)

Website Type	Purpose	Example & Founding Year
Social Networking	Users can interact with friends and followers and develop new social connections.	Myspace (2003) Facebook (2004) Twitter (2006)
Social Review	Customers leave reviews for other potential customers.	TripAdvisor (2000) Yelp (2004)
Image Sharing	Users can share visual content with other users and followers.	Instagram (2010) Pinterest (2010) Snapchat (2011) VSCO (2011)
Video Hosting	Users can create and share videos for streaming by other users.	Vimeo (2004) YouTube (2005)
Community Blog	Users can express their thoughts and ideas in a blog format.	Tumblr (2007) Medium (2012)
Discussion	Specifically designed to facilitate discussions among users.	Reddit (2005) Quora (2009)
Sharing Economy Networks	Allows users to share resources with one another.	Airbnb (2008) Rover (2011)
Source: https://biteable.com/blog/tips/the-7-different-types-of-social-media/		

allow users to both create and share videos for others to stream online. Community blogs allow users to express their feelings on various topics, while discussion websites, such as Reddit and Quora help facilitate conversations and debates. Finally, sharing economy networks provide peer-to-peer selling and sharing of resources based on an individual's specific needs (e.g., a dog walker, overnight accommodations for a trip, etc.) It is important to note that there is connectivity between many of the social media platforms. For example, a user can post an image on Instagram and then share that post with friends and followers on Facebook.

How can Social Media be used to Facilitate Social Movements?

Social media has become influential in promoting and organizing social movements because it combines the use of digital communication and engaging dialogues to solve social problems. With the use of social networking sites, anyone can be aware of a global or local event within seconds. As videos of current events are posted online, testimonies often flood news feeds and there is a personal touch added to each event. Once there is awareness of an issue, social media facilitates a place to organize the cause and show support to others. Gerbaudo (2012) argues that the media provides a channel through which social movements are organized and further developed. Access to social media is a crucial tool for large groups to organize. The information is disbursed in a timely manner to a large audience. Developments are made by contributing ideas to the future of the movement. Being online allows everyone's voices to be heard. In place of traditional media's designated spokespersons, multiple voices from all walks of life produce user-generated content in social media (Freberg, 2013).

In the digital era, there is a heightened awareness about social issues. Accountability for being aware of the discourse that occurs in our communities is important. Social media allows individuals to participate in activism and demonstrate support to causes an individual may not be able to reach physically. The use of hashtags makes the process of claim-making easier because it gains momentum as it is used at a large rate. Previously mentioned social media sites, Facebook, Twitter, and Instagram have trending pages for top mentioned topics and hashtags. For example, the #IceBucketChallenge brought awareness to the ALS (Amyotrophic Lateral Sclerosis, also known as Lou Gehrig's disease) Association. What started as a fun way to challenge others to donate money raised \$115 million in the summer of 2014 (Rogers, 2016). In 2018 the #metoo hashtag demonstrated support for survivors of sexual harassment and assault, which brought increased awareness to the pervasiveness of these crimes (Johnson & Hawbaker, 2018).

#BlackLivesMatter

The BLM hashtag serves as the *coalescence* stage of this social movement; it brought awareness to social injustice while bringing together a community against police brutality. The BLM movement was started by three Black women; Alicia Garza, Patrisse Cullors, and Opal Tometi (Garza, 2014). These women came together to oppose police violence and injustice against Blacks, later to expand their cause. The hashtag #Blacklivesmatter was created in July 2013 in response to George Zimmerman's acquittal for the second-degree murder of an African American teenager, Trayvon Martin (Freelon, McIlwain, & Clark, 2016). After July 2013 through May 1, 2018, the hashtag has been used nearly 30 million times on Twitter (Anderson, Toor, Rainie, & Smith, 2018). After gaining a core group of participants, the founders published their list of demands that would be the guiding principles of the BLM association. According to D.C. Area Educators for Social Justice: A Project for Teaching Change (2018), included in the guiding principles are restorative justice, diversity, queer-affirming, globalism, unapologetically

Black, collective value, empathy, loving engagement, transgender affirming, Black villages, Black women, Black families, and intergenerational.

After June 2014, the hashtag use was declining on the Twitter platform. Months later after the death of Eric Garner, there was an uptick followed by silences. Twitter activity around this time was still relatively small in scale compared to the attention to come after the death of Mike Brown (Freelon, McIlwain, & Clark, 2016). The death of Mike Brown in Ferguson, Missouri was the catalyst that made BLM the well-known political activist group today. After June 2014, the hashtag use was declining on Twitter (Anderson, Toor, Rainie, and Smith 2018). Over the past five years BLM has been used on a continuous pace with significant events. According to findings from the Pew Research Center, the BLM hashtag was mentioned an average of nearly 500,000 tweets daily between July 7th and 17th (Anderson et al., 2018). This social movement has surpassed the hashtag, and according to Lebron (2017), BLM is a powerful force demanding social change in America. In the case of BLM, social media was a symbolic megaphone, allowing voices of neglect, mistreatment, and injustice to share their stories for the world to hear.

THEORETICAL CONSIDERATIONS

The application of sociological theory can provide a better understanding of what social media means for modern social change. In order to explain collective behavior, such as social movements, some scholars argue that it is imperative to pay attention to resources, and the *resource mobilization perspective* suggests that the aggregation of resources, that is, money and labor are critical for an understanding of social movement activity (McCarthy & Zald, 1997). Resource mobilization theory emphasizes resources, organization, and political opportunities as a means for social change (Jenkins, 1983). Resource mobilization theorists suggest that social movements are extensions of institutionalized actions, focusing on institutional change and previously unorganized groups that form together against institutional elites (Jenkins, 1983).

The social psychological version of resource mobilization theory suggests that at the individual-level, a person weighs the costs and benefits of participating in a social movement and (1) knows the opportunities to participate, (2) determines if he/she is capable of engaging in the opportunity, and (3) decides if he/she is willing to participant (Klandermans, 1984). More recently, Polletta and Jasper (2001) focus on collective identities and why they matter to social movements. They argue that the term *collective identity* describes both imagined and concrete communities that are fluid and relational. Collective identities form through bystanders, allies, opponents, and news media (Polletta & Jasper, 2001). They also argue that collective identities matter to social movements because they promote new identities as a mechanism to achieve power and transform selves (Polletta & Jasper, 2001).

Drawing from resource mobilization theory, it is important to consider that social media acting as an avenue for social change is limited by technological capital. Individuals' access to the Internet and social media is based on access to and the availability of valued resources, which depends on the demographic characteristics and resources of particular groups (Carlson & Isaacs, 2018). That is, not everyone in society has equal access to different forms of technology, including the Internet. This means that social media may be more influential in mobilizing certain sectors of the population, including those with the greatest resources. While researchers suggest that the number of people *not* engaged with social media are dwindling, there are still those without it, which imposes a limit for the use of social media as a platform for facilitating social change (Anderson & Hitlin, 2016). Because access to technology is

a privilege not everyone has, online protests may be excluding some of the most marginalized groups in society. Underrepresented groups in America often lack a voice in matters of social, political, and economic injustice, and they are the least likely to have access to technology. This in turn, restricts their participation in social media and online social movements.

To forge a multi-dimensional understanding of the struggles of some of the most marginalized groups in society, the theoretical framework, intersectionality focuses on multiple social identities of individuals and groups and society, including race, social class, gender, sexuality, disability, nationality, among others. Drawing from Crenshaw's (1989) *intersectionality*, Collins (2000) argued that an intersectional approach is particularly useful in order to gain a more thorough understanding regarding how forms of intersecting categories of oppression influence individuals' lives. An intersectional approach underscores that discrimination and inequality are experienced differently among marginalized groups, which is relevant to the manner in which people individually and collectively organize to fight against injustice. Individuals do not experience injustice as one category of their social identity. Instead, the ways in which gender, race, sexuality and other social identities are entangled, shapes individuals' experiences with various forms of injustice in the United States (Collins 2000). Chun, Lipsitz, and Shin (2013) argued that intersectionality infuses collective behaviors, especially those organized by women of color. As we see with BLM, the original social movement was started by three individuals who identified in at least two marginalized groups: Black women. Therefore, intersectionality is useful in demonstrating that discrimination and inequality are experienced differently not only among marginalized groups, but also among groups of individuals experiencing multiple forms of oppression (Choo & Ferree, 2010; Collins, 2000; Miller, 2006; Moore, 2011).

FUTURE RESEARCH

Social media is still a relatively new phenomenon and because of this, prior research is limited. Longitudinal research is extremely beneficial in studying social changes long-term because of the opportunity to see changes in habits and behaviors over time. Within the realm of social media, longitudinal research is lacking because of its novelty, meaning there lacks a significant amount of information for data comparison. Previously conducted studies have been restricted to individual social media platforms such as Twitter or Facebook (Tow et al., 2010; Khan & Jarvenpaa, 2010). Considering recent and ongoing technological advances, integration of social media platforms now allows for easier and more complete research across online sites. Additionally, age is a limiting factor in prior research. At its conception, social media was associated and marketing to younger users, and scholars have largely failed to consider the older population use of social media sites. Consequently, there are gaps in the research regarding the population as a whole and the link between social media and social change (Vishwanath, 2015).

In addition to considering the use of longitudinal data collection and analysis, drawing from Harlow's (2011) research on social media and social movements, the authors of this article recommend that future research should consider how social movements may begin online and then move offline or vice versa. Internet users may become interested in a topic from a post or hashtag on a social media site, which may in turn, may contribute to offline protests and further collective action (Harlow, 2011). Or offline grassroots organizations may gain social media attention, which gives their collective action further momentum, facilitating it into further stages of a social movement. Future research should consider these unique paths as organizations fight for justice, including when social media is a driving force behind a

social movement or when social media advances an already present social movement further igniting pro-justice collective action.

CONCLUSION

The ability of individuals to come together to advocate for social change is an important part of America's history and tradition. In fact, over the past several decades, several social movements have transformed the culture of American society, including the Labor Movement, the Women's Movement, the Civil Rights Movement, and the LGBT Movement. More recently, Occupy Wall Street brought awareness to income inequality and inequity in American society, and today, we are seeing the power of the Internet as the BLM hashtag has become a key symbol of a social movement against police brutality, violence, and racism against African Americans. As social media websites and platforms continue to develop and change, we may see a renewal of prior social movements, and the growth of social awareness about new social causes. It is important that all people have a voice in advocating for change; consequently, the Internet allows for each person to have a voice. Through Facebook, Twitter, Instagram and other social media websites, the ability to connect and network with others is what helps propel that one voice into a larger group of voices who together, have the power to fight for change and justice.

REFERENCES

- Anderson, M., & Hitlin, P. (2016). Social media conversations about race: How social media users see, share and discuss race and the rise of hashtags like #BlackLivesMatter. *Pew Research Center*. Retrieved from http://www.pewinternet.org/2016/08/15/social-media-conversations-about-race/pi_2016-08-15_race-and-social-media_3-01/
- Anderson, M., Toor, S., Rainie, L., & Smith, A. (2018). An analysis of #BlackLivesMatter and other Twitter hashtags related to political or social issues. *Pew Research Center*. Retrieved from <http://www.pewinternet.org/2018/07/11/an-analysis-of-blacklivesmatter-and-other-twitter-hashtags-related-to-political-or-social-issues/>
- Beal, V. (2019). Follower. *Webopedia*. Retrieved from <https://www.webopedia.com/TERM/F/follower.html>
- Browder v. Gayle, 352 U.S. 903. (1956). *Stanford, The Martin Luther King, Jr. Research and Education Institute*. Retrieved from <https://kinginstitute.stanford.edu/encyclopedia/browder-v-gayle-352-us-903>
- Carlson, A., & Isaacs, A. M. (2018). Technological capital: An alternative to the digital divide. *Journal of Applied Communication Research*, 46(2), 243–265. doi:10.1080/00909882.2018.1437279
- Carroll, W. K., & Hackett, R. A. (2006). Democratic media activism through the lens of social movement theory. *Media Culture & Society*, 28(1), 83–104. doi:10.1177/0163443706059289
- Choo, H. Y., & Ferree, M. M. (2010). Practicing intersectionality in sociological research: A critical analysis of inclusions, interactions, and institutions in the study of inequalities. *Sociological Theory*, 28(2), 129–149. doi:10.1111/j.1467-9558.2010.01370.x
- Chun, J. J., Lipsitz, G., & Shin, Y. (2013). Intersectionality as a social movement strategy: Asian immigrant women advocates. *Signs (Chicago, Ill.)*, 38(4), 917–940. doi:10.1086/669575

Collins, P. H. (2000). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *University of Chicago Legal Forum*, 14, 538–554.

D.C. Area Educators for Social Justice. A Project for Teaching Change. (2018). *Black lives matter week of action-FAQ*. Retrieved from <https://www.dcareaducators4socialjustice.org/black-lives-matter-week-faq/>

Freberg, K. (2013). Social Media. In R. L. Heath (Ed.), *Encyclopedia of public relations* (2nd ed.). Sage Publications. doi:10.4135/9781452276236.n457

Freelon, D., McIlwain, C. D., & Clark, M. D. (2016). Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. *Center for Media & social Impact*. Retrieved from http://cmsimpact.org/wp-content/uploads/2016/03/beyond_the_hashtags_2016.pdf

Friedland, L. A. (2013). Social Media. In V. Smith (Ed.), *Sociology of work: An encyclopedia*. Sage Publications. doi:10.4135/9781452276199.n284

Garrett, K. R. (2006). Protest in an information society: A review of literature on social movements and new ICTs. *Information Communication and Society*, 9(02), 202–224. doi:10.1080/13691180600630773

Garza, A. (2014). A herstory of the #BlackLivesMatter movement. In J. Hobson (Ed.), *Are all the women still white* (pp. 23–28). State University of New York Press.

Gerbaudo, P. (2012). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

Goodwin, J., & Jasper, J. M. (2009). *The social movements reader: Cases and concepts* (2nd ed.). Wiley-Blackwell.

Harlow, S. (2012). Social media and social movements: Facebook and an online Guatemalan justice movement that moved offline. *New Media & Society*, 14(2), 225–243. doi:10.1177/1461444811410408

Jenkins, J. C. (1983). Resource mobilization theory and the study of social movements. *Annual Review of Sociology*, 9(1), 527–553. doi:10.1146/annurev.so.09.080183.002523

Johnson, C. A., & Hawbaker, K. T. (2018). #MeToo: A timeline of events. *Chicago Tribune*. Retrieved from <http://www.chicagotribune.com/lifestyles/ct-me-too-timeline-20171208-htmlstory.html>

Jones-Eversley, S., Adedoyin, A. C., Robinson, M. A., & Moore, S. E. (2017). Protesting Black Inequality: A Commentary on the Civil Rights Movement and Black Lives Matter. *Journal of Community Practice*, 25(3–4), 309–324. doi:10.1080/10705422.2017.1367343

Khan, Z., & Jarvenpaa, S. L. (2010). Exploring temporal coordination of events with Facebook. *Com. Journal of Information Technology*, 25(2), 137–151. doi:10.1057/jit.2010.8

Klandermans, B. (1984). Mobilization and participation: Social-psychological expansions of resource mobilization theory. *American Sociological Review*, 49(5), 583–600. doi:10.2307/2095417

Lebron, C. J. (2017). *The making of black lives matter: A brief history of an idea*. Oxford University Press.

- McCarthy, J. D., & Zald, M. N. (1977). Resource mobilization and social movements: A partial theory. *American Journal of Sociology*, 82(6), 1212–1241. doi:10.1086/226464
- Miller, J. (2006). *Getting played: African American girls, urban inequality, and gendered violence*. New York University Press.
- Moore, M. R. (2011). *Invisible families: Gay identities, relationships, and motherhood among Black women*. University of California Press.
- Morozov, E. (2011). *The net delusion: The dark side of internet freedom*. Penguin.
- Polletta, F. (2006). *It was like a fever: Storytelling in protest and politics*. University of Chicago Press. doi:10.7208/chicago/9780226673776.001.0001
- Polletta, F., & Jasper, J. M. (2001). Collective identity and social movements. *Annual Review of Sociology*, 27(1), 283–305. doi:10.1146/annurev.soc.27.1.283
- Rogers, K. (2016). The “ice bucket challenge” helped scientists discover a new gene tied to A.L.S. *The New York Times*. Retrieved from <https://www.nytimes.com/2016/07/28/health/the-ice-bucket-challenge-helped-scientists-discover-a-new-gene-tied-to-als.html>
- Statista. (2019). <https://www.statista.com/>
- Tow, W. N. F. H., Dell, P., & Venable, J. (2010). Understanding information disclosure behaviour in Australian Facebook users. *Journal of Information Technology*, 25(2), 126–136. doi:10.1057/jit.2010.18
- Vishwanath, A. (2015). Diffusion of deception in social media: Social contagion effects and its antecedents. *Information Systems Frontiers*, 17(6), 1353–1367. doi:10.1007/10796-014-9509-2

ADDITIONAL READING

- Bonilla, Y., & Rosa, J. (2015). # Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist*, 42(1), 4–17. doi:10.1111/amet.12112
- Freelon, D., McIlwain, C., & Clark, M. (2018). Quantifying the power and consequences of social media protest. *New Media & Society*, 20(3), 990–1011. doi:10.1177/1461444816676646
- González-Bailón, S., & Wang, N. (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks*, 44, 95–104. doi:10.1016/j.socnet.2015.07.003
- Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J., & Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology*, 39, 85–118. doi:10.1111/pops.12478
- LeFebvre, R. K., & Armstrong, C. (2018). Grievance-based social movement mobilization in the # Ferguson Twitter storm. *New Media & Society*, 20(1), 8–28. doi:10.1177/1461444816644697
- Uldam, J. (2018). Social media visibility: Challenges to activism. *Media Culture & Society*, 40(1), 41–58. doi:10.1177/0163443717704997

KEY TERMS AND DEFINITIONS

8

#BlackLivesMatter (BLM): A hashtag, created in 2013 that campaigns against police brutality, other forms of violence, and systemic racism against Black people.

Follower: A term used to describe an individual who subscribes to receive a social media user's updates on social media websites, including Facebook, Instagram, and Twitter.

Intersectionality: With an emphasis on the experiences of women of color, intersectionality is a theoretical perspective, developed by Kimberlé Crenshaw in 1989 that focuses on marginalized individuals' intersecting social identities.

Resource Mobilization Theory: A theoretical perspective that addresses the resources (time and money), organization, and political opportunities needed for social change.

Social Media: Online platforms that are used for social networking and sharing information and ideas.

Social Movement: Individuals or groups of individuals that come together to facilitate social change.

Digital Social Networks From a Social Capital Perspective

Suparna Dhar

NSHM Knowledge Campus, Kolkata, India

Indranil Bose

Indian Institute of Management, Kolkata, India

Mohammed Naved Khan

Aligarh Muslim University, India

INTRODUCTION

Digital Social Networking (DSN) sites such as Facebook, Twitter, LinkedIn, WhatsApp, Instagram, Pinterest among many others have garnered millions of users worldwide. This instance of information and communication technology has brought about changes in the way people communicate and interact affecting human lifestyle across the world. According to many, the phenomenon has permeated geographical and cultural boundaries, disregarded age and gender barriers, diffused social and economic strata, and transgressed linguistic and ethnic boundaries. Such widespread and intense impact of technology on human life is unprecedented in human history. The unique opportunities offered by digitization of social networking, the novel capabilities of information dissemination and foraging on DSN sites, along with widespread adoption have generated academic interest in the theoretical interpretation of antecedents and consequences of DSN adoption. In this chapter, the authors offer a social capital perspective of DSN adoption.

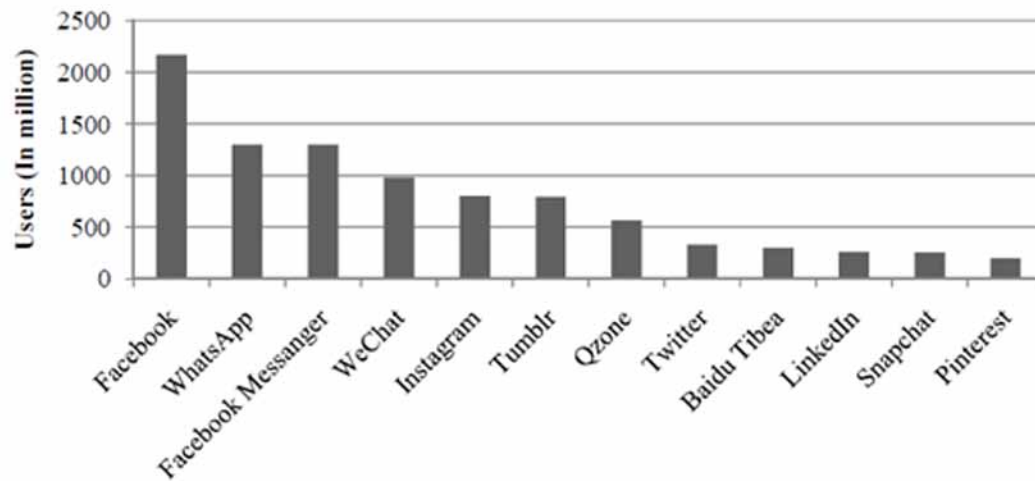
Social capital is a manifestation of the underlying social network. It has a symbiotic relationship with social network. Social capital is determined by network mechanism, and reciprocity, both at a general and individual level. The concept of social capital applies the economic principle of rational action and economic systems to social systems maintaining its social structure. Like human capital, social capital is less tangible than physical capital but displays the potential to facilitate economic output by boosting productivity and trust. Human capital is an individual attribute, whereas social capital is an attribute of human relationships. Social capital allows conversion to other forms of capital such as human capital and intellectual capital. It distinguishes itself from financial and human capital as it is owned wholly or in parts by individuals and acts as a catalyst in mobilizing financial and human capital.

BACKGROUND

DSN is a culmination of the rapid evolution of information technology, aided by revolutionary advancements in telecom and computing technologies. The first decade of the twenty-first century witnessed the emergence of DSN sites such as Friendster, MySpace, LinkedIn, Flickr, Facebook, Twitter, WhatsApp, Instagram and many others, refer Figure 1. Millennials have adopted DSN as part of their daily

DOI: 10.4018/978-1-7998-3479-3.ch076

Figure 1. Users on popular DSN sites (adapted from Statista, 2018)



existence. It has empowered users to create and share content, a big difference from general users being mere consumers of content. The rapid proliferation and wide adoption of DSN has had a profound effect on social, economic, political, organizational and emotional aspects of human life (Kane et al., 2014).

DSN sites are defined as “web based services that enable users to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system” (boyd & Ellison, 2007, p. 2). DSN allows users to create and exchange user-generated content, which empowers general users to publicly voice their opinions and views and find an audience for their creativity. DSN sites constitute websites facilitating networking or relationship building between people. Networking typically refers to forming a relationship with strangers. DSN allows people to network - to connect with strangers and expand their social horizon. It allows users to articulate their profile and make their profile and connections visible in the network (boyd & Ellison, 2007). The profile search feature on SNS aids users find people with similar interests and form new connections. It allows users maintain the relationship with their existing networks. In consequence, it aids users to maintain a larger network on DSN compared to offline social networks.

DSN affordances of expanded social reach aided by communication richness grants a paradigm shift towards enhanced social networking. This endows DSN users an opportunity to enhance their social capital on DSN. Uslaner (2000) proposed that spending more time on the internet encumbers face to face interaction which is detrimental to social relationships. DSN has belied this dystopian view - it promotes social interaction. It facilitates the meeting of like-minded people forming communities of common interest and shared goals. This improved interaction promotes social capital (Hofer & Aubert, 2013). Academic studies found that DSN show positive association with community interaction, involvement, and social capital (Ellison et al., 2007; Ellison et al., 2011; Hampton & Wellman, 2003).

SOCIAL NETWORKING

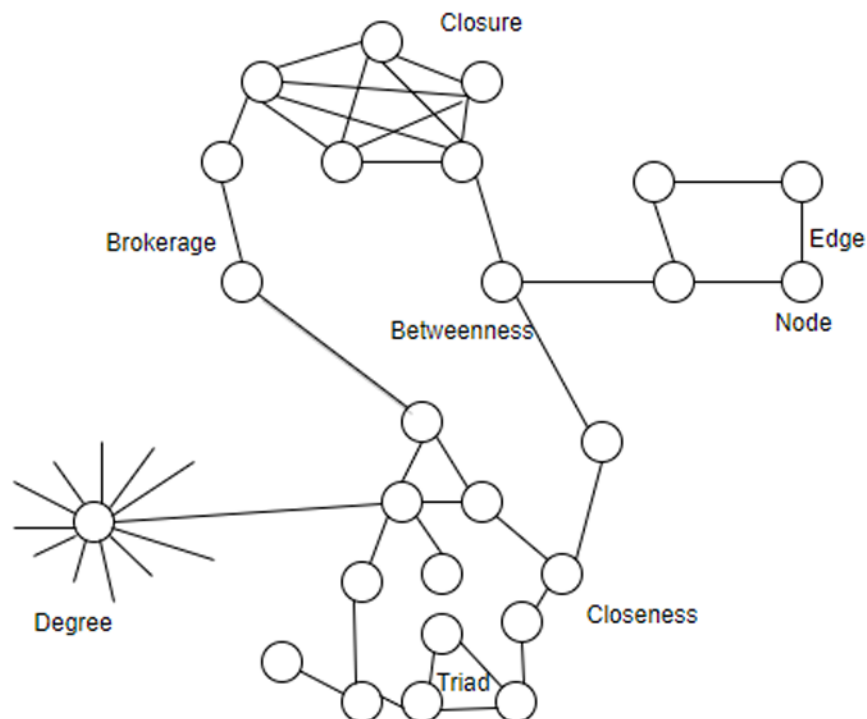
Social network theory, proposed by Jacob (1934), interprets social relations and actions for a contextual, relational and systemic understanding of human behavior. In social network theory, social relations are

represented by a structure of nodes and edges, where the nodes represent people or groups, and the edges represent interaction, collaboration and influence between the nodes. A number of nodes and edges together represent the social network. Networks constitute direct and indirect ties between nodes. Direct ties between nodes form relationship dyads and triads. Structural holes indicate absence of ties for triadic closure (Bapna et al., 2017). Network analysis shows that nodes associated with large number of structural holes enjoy more power in the network (Burt, 1992a). Directed network ties (example: manager to a subordinate) comprise of unequally connected nodes that support unidirectional or unbalanced network flow. Undirected network ties (example: peers) constitute nodes with equal connection on either side where network flows are bidirectional. Different types of ties represent different types of relations (example: partner, competitor).

In the social network, the tie-strength measures the strength of the relationship (Granovetter, 1973). Weak ties constitute loose connections between people that serve as information conduit (Putnam, 2000). Weak ties in a network are inclusive in nature and effective in broadening one's social horizons, connecting with people from different life situations and opening up opportunities for more information and resources (Williams, 2006). Strong ties are seen between close friends and family members. The term ego-network (Koka & Prescott, 2001) refers to a network represented by ties of a focal node, where the focal node is called an ego, and the other nodes are known as alters. In a larger context, academicians refer to the network structure as a network of networks, where each individual network displays stronger tie strength between its nodes and is distinguished by some intrinsic characteristics; and yet the smaller networks are not discrete, there are interlinks and overlaps.

Academic researchers have characterized network structures, refer Figure 2, by centrality (the degree to which relations are guided by network hierarchy), closeness (the network distance), betweenness (node

Figure 2. Social network



lies between other nodes in the network), multiplexity (multiple network paths between nodes), holes (nodes is connected to two others which are not mutually connected) and cliques (clusters of networks connected to one another) attributes (Liben-Nowell & Kleinberg, 2007). The attributes of the nodes and the attributes of the ties produce a distinct effect on the influence of the tie. Attribute-based differentiation generates cliques or clusters within the network, such as location-based clique or skill based clique. Social network theorists have posited that the network influences are the effect of ties and network position, decimating the need to study attributes of the node. This view theorizes that the attributes of the nodes manifest in their position, as similar nodes occupy similar positions in the network (Wellman et al., 1996). Rowley's (1997) social network model studied the network density and organization's centrality in the network to examine structural influences. In network terms, density refers to the relative number of ties linking network nodes. Centrality of network nodes determines the influence of the node to group proficiency and behavior (Kane & Borgatti, 2011).

Social network researchers study the formation, expansion and changes in social structures, impact of characteristics of the network, likelihood of ties between members of the network and formation of social capital (Pool & Kochen, 1979). Structure of the social networks and allocation and flow of resources within the network constitute predicting factors of social behavior in the network (Wellman et al., 1996).

DIGITAL SOCIAL NETWORK

DSN has gained wide popularity across the globe. It has become the de facto platform for users from varied backgrounds to connect with friends; share and find content. DSN allows users to create personal profiles, search profiles in the network, form network, create communities, generate content and share information and views which are easily dispersed and discovered in the digital network. The backbone of digital social networking sites consist of profiles having descriptors like age, location, interests, etc. and an articulated list of friends or connections. Information shared on these sites circulates across the globe instantly and generates comments and unique responses such as "like" and "share" (Lipsman et al., 2012). Users search SNS profiles to locate offline friends and form new connections, often with like-minded people evincing similar interests and views (Ellison et al., 2007). Table 1 lists typical features of DSN.

Prevalent DNS sites present a fairly consistent set of key network characteristics as well as technological features. Some minor variations in implementation of social networking characteristics along with variance in focus result in attracting different types of users and user behaviors (Lee & Suh, 2013). The sites vary in the extent of profile visibility, and access to connections and profile content. For example, Facebook requires elaboration of user's personal information for profile creation; Twitter preserves users' anonymity, thus creating a structural difference between the two platforms. User anonymity allows

Table 1. Typical features of digital social networking platforms

Feature	Description
Digital profile	Profiles having descriptors like age, location, interests and an articulated list of friends or connections.
Search	Search for profiles and contents and traverse the network.
Privacy	Allow or limit access to profile and digital content.
Content generation	Generate textual and multimedia content and share information and views.
Digital responses	Responses include text, "like", "share", emoticons.

focus on the communication content as opposed to the perpetrator of the information. Facebook supports bidirectional connections where the connection represents friendship that manifests equally on both the nodes. Twitter supports directed connections, with a follower on one end of the connection and a followed on the other end (Hofer & Aubert, 2013). Network closeness showed positive association between self-disclosure and communication frequency on Facebook (Bapna et al., 2017). These implementation variations lead to variation in performance of DSN sites and the consequences of DSN use.

DIGITAL VERSUS TRADITIONAL SOCIAL NETWORKING

DSN allows users to voice their opinions, views and share content and percolates the user generated content to a large audience through the network (Kaplan & Haenlein, 2010). Digital personal profiles are articulated and discoverable in the network. Profile and content search feature of online SNS platforms allow users to locate profiles with specific attributes. DSN is dynamic, growing and change rapidly over time, with the addition of nodes and new interchanges along the edges (Liben-Nowell & Kleinberg, 2007). It facilitates relatively more uninhibited and creative communication compared to in-person offline communication. Online social networking platforms allow systematic and quick flow of information and social support, which differ from the content flow in traditional social networks (Kane et al., 2014). The platform offers novel affordances, such as content multiplicity and unique responses such as “like”, which are not supported in traditional offline communications (Lipsman et al., 2012). The communication, interaction and collaboration capabilities of online SNS outstrip time and space limitations. This small world has become even smaller with DSN (Fu, Liu, & Wang, 2008). Offline social networks are mired with network cliques that displayed intense relationships in the inner circle (Dunbar & Spoor, 1995). DSN allows formation of new ties to expand the network boundary and form inter-group ties garnering more openness and inclusiveness in the network. Social stratification of like individuals increases the probability of two individuals having a minimum referral chain with a common friend, but the awareness of this minimum chain is uncertain in offline scenario (Pool & Kochen, 1979). Identifying the right path in the chain is difficult in offline social networks. Online referral webs have eased identification of minimum referral chains (Kautz et al., 1997).

DSN sites have increased the availability of data on social networks, which improved opportunities to study the basic properties and recurring structural features of social networks (Liben-Nowell & Kleinberg, 2007). It offers an easy and economical means to form and maintain a large number of ties, thus allowing users nurture a larger number of existing offline ties as well as new ties formed online (Ellison et al., 2007) to expand their network. Online communities have allowed individuals to shift from location-based networks or offline social groups to interest-based networks (Wellman, et al., 1996). Economic access to DSN has promoted subscription to interest-based virtual communities (Williams, 2006). On DSN platforms, ties are more observable, searchable and quantifiable compared to traditional social networks. This has improved opportunities for social interaction and consequent social capital generation.

SOCIAL CAPITAL

The term social capital (refer Table 2) appeared in social science literature dating back to the early twentieth century (Hofer & Aubert, 2013). The concept resurfaced in the 1960s (Putnam, 2000). The concept of social capital applies the economics principle of rational action (Coleman, 1988). Like human

Table 2. Social capital definitions

Author	Definition
Hanifan (1916)	"... good-will, fellowship, mutual sympathy and social intercourse among a group of individuals and families who make up a social unit" (p. 130)
Coleman (1988)	"Social capital is defined by its function. It is not a single entity but a variety of different entities, with two elements in common: they all consist of some aspect of social structure, and they facilitate certain actions of actors – whether persons or corporate actors - within the structure" (p. 302)
Burt (1992b)	"...resources available to any one person in a population are contingent on the resources available to individuals socially proximate to the person." (p. 60)
Adler and Kwon (2002)	"Social capital-understood roughly as the good-will that is engendered by the fabric of social relations and that can be mobilized to facilitate action..." (p. 17)

capital (Becker, 1962), social capital is less tangible than physical capital. It is an attribute of human relationships (Woolcock, 1998). Social capital acts as a catalyst in mobilizing resources and support to diffuse financial and human capital (Burt, 1992a). It yields individual and organizational performance variations by driving coordination, creativity, leadership, learning, teamwork, trust, career success, reduced turnover rate (Adler & Kwon, 2002).

Social capital is typically measured by tie strength (Helliwell & Putnam, 2004). Weak ties constitute loose connections between people. These ties serve as information conduit (Putnam, 2000). Putnam (2000) coined the terms "bridging" and "bonding" social capital (Williams, 2006). Weak ties in a network lead to bridging social capital, which is inclusive in nature and effective in broadening one's social horizons, connecting with people from different life situations and opening up opportunities for more information and resources (Williams, 2006). Strong ties, as seen between close friends and family members, lead to bonding social capital (Williams, 2006).

Past researchers studied social capital formation in terms of psychological imports such as self-efficacy, well-being, emotional support; social imports such as solidarity, trust and trustworthiness, civic and political impacts; and business gain (Burt, 1992b; Ellison et al., 2007; Granovetter, 1973). Increase in social capital leads to increase in commitment to the community, thus yielding benefits. Individuals leverage social capital to draw on resources from other members of the network. Declining social capital reduce trust among community members. Structural dimension of social network deals with social capital generated by virtue of the network, network structure, ties, tie strength, positions. Relational dimension deals with relationship attributes such as trust and goodwill. Cognitive dimension involves shared goals and norms. The cohesive strength of a community and collective efficacy determine social capital. Gaining access to resources in the social network is the primary function of social capital theory. The structural, relational and cognitive dimensions collectively influence mobilization of resources in the social network.

ROLE OF DIGITAL SOCIAL NETWORKING IN SOCIAL CAPITAL FORMATION

Researchers have often debated whether internet use leads to increase or decrease in social capital. When people spend more time on the internet, it results in less time and intent to meet people face to face. Consequently, stronger engagement with online platforms leads to a poor offline interpersonal relationship. On the other hand, enhanced interaction and larger number of connections supported by digital social networking is expected to facilitate formation of social capital. DSN allows users to maintain

touch with past ties leading to maintained social capital (Ellison et al., 2007). Frequent interactions in the network on diverse topics of common interests help users strengthen their social relationship ties, which induce relatively weaker ties to beget more bonding form of social capital. Information disclosure on DSN enables extending the network of weak ties, which in turn mobilizes resources and information (Ellison, Steinfield, & Lampe, 2011). Since DSN sites provision maintenance of both strong and weak ties, it serves as a platform for maintaining both bonding and bridging social capital. On the flipside, DSN allows connection across social, geographical and cultural boundaries, introducing complex virtual network relationships that could occasionally turn out to be dangerous and harmful.

SNS users maintain a connection with offline friends and form new connections online to form weak ties (Ellison et al., 2007). Bridging social capital emanates from weak ties or loose connections between people, which serves as an information conduit (Putnam, 2000). Ellison et al. (2007) study found a strong association between Facebook use and all types of social capital, where the association between Facebook use and bridging social capital was the strongest. Hofer and Aubert (2013) showed a differential effect of the directed network on social capital formation in their study of social capital generation on Twitter. Interestingly, their study showed that the number of followers was associated with bonding social capital and number of followees was associated with bridging social capital (Williams, 2006). Morris et al. (2010) found trusted information exchanges to be a motivation for information seeking on DSN. Wider and more frequent interactions on DSN serve to strengthen ties with network connections. The stronger network ties result in higher reciprocation and solidarity, emanating emotional support, and social and tangible benefits, which were determinants of bonding social capital. Ellison et al. (2011) found DSN use to exhibit significant association with bonding social capital.

IMPLICATIONS OF SOCIAL CAPITAL

Increase in social capital leads to increase in commitment to the community, thus yielding benefits. For example, social capital yields individual and organizational performance variations by driving coordination, creativity, leadership, learning, teamwork, trust, etc. Social capital facilitates psychological imports such as self-efficacy, well-being, and emotional support. The cognitive dimension of social capital capacitates collective goals to influence individual and group behavior, which improves group performance. It mobilizes workforce to adhere to quality processes, value-add and achieve continuous improvement.

Social capital is a core concept in business. Organizations gain competitiveness and generate returns from investments in social capital (Burt, 1992b), similar to investments in financial and human capital. Bridging social capital is effective in case of job searching, information diffusion, product quality, knowledge management and firm competitiveness (Borgatti & Foster, 2003; Granovetter, 1973; Putnam, 2000). Bonding social capital provides more substantive support that is useful in organization transformation (Williams, 2006). In large firms, sharing visions and values and understanding of common goals are necessary to develop the firm's social capital, useful in forming strategic alliances and buyer-supplier networks (Koka & Prescott, 2001). Social capital eases knowledge capture and its flow in large multinational corporations working with a global network of suppliers and other stakeholders. Information and resource access in the network through social capital is useful in career success of professionals (Seibert, Kraimer, & Liden, 2001).

Social capital in the organization's employee network generates job satisfaction, which fosters voluntary actions, value and innovation and promotes entrepreneurship (Koka & Prescott, 2001). In an external organizational network, social capital with strategic industry leaders, regulators and other people hav-

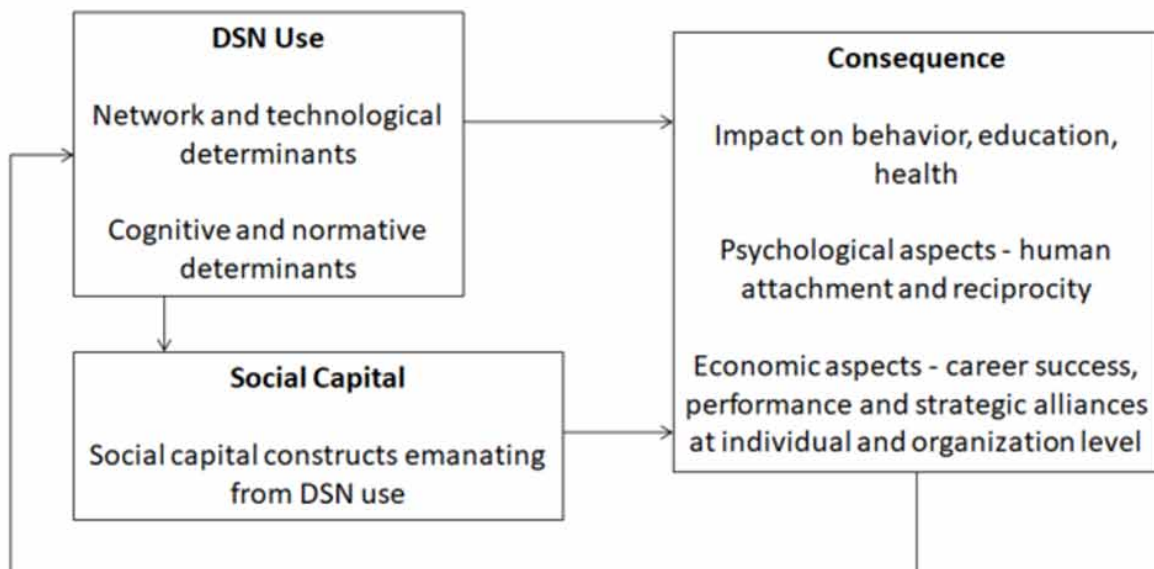
ing the power to influence organization performance provides a competitive advantage. Social capital in competitor network offers opportunities for beneficial collaborations. In short, use of social capital endows an organization the ability to maintain a strategic competitive advantage (Tsai, 2000).

Independent empirical research conducted by the authors of this chapter reinforced generation of social capital formed on DSN (Dhar, 2018). Principal component analysis on social capital items in the research instrument yielded three factors, where the bridging social capital sustained as a single factor, but the bonding social capital yielded two independent factors, namely, resource access and social support. This research showed positive association between DSN use and generation of all social capital factors. The study indicated perceived business benefits on strategic and operational dimensions from DSN use partially mediated by social capital formed on DSN. Thus, from a utilitarian perspective, social capital constitutes a critical component in determining consequences of DSN use.

FUTURE RESEARCH DIRECTIONS

Social capital is increasingly gaining importance in academic literature for its economic imports. DSN such as Facebook and Twitter allows unique social networking capabilities in terms of network structure and information flow in the network. This introduces new dimensions to social capital generated on DSN. Social capital influences a multitude of social aspects such as behavior, education, health and economic development (Adler & Kwon, 2002). It affects psychological aspects such as human attachment and reciprocity (Wasko & Faraj, 2005). It also affects economic aspects such as career success, performance and strategic alliances at individual and organization level (Seibert et al., 2001; Tsai, 2000). In consequence, differentiated social capital formation on DSN affects social, psychological and economic aspects of human life. There exists gap in academic research on understanding imports of DSN (Kane et al., 2014). The social capital perspective of DSN use portrays diverse potential consequences of DSN that call for research (refer Figure 3).

Figure 3. Future Research Agenda



Future researchers need to study the network and technological determinants of DSN that influence social capital formation. For example, novel technology attributes such as “like” manifests unique emotional consequences (Lipsman et al., 2012). Such emotions influence social interactions and relationships leading to differences in social bonding and intensity of social capital. Secondly, researchers need to determine social capital constructs emanating from DSN use. In a study conducted by Ellison et al. (2007) on association between DSN intensity and social capital, bonding capital items yielded two unique factors on, while the bridging social capital sustained as a single factor. This indicates the possibility of DSN use yielding new dimensions of social capital. Thirdly, researchers need to investigate the specific consequences of DSN use and the mediation role of social capital in the relationship. For example, DSN allows traversing network connections to connect with profiles having common interests, leading to bridging social capital. This is useful in knowledge sharing in distributed organizations (Ellison et al., 2015). Researchers need to explore similar consequences in organizational business functions, education, healthcare, old age care and other economic imports.

CONCLUSION

DSN is perhaps one of the greatest social upheavals of the twenty-first century. DSN adoption and influence have surpassed all past technology implementations. There exists a debate on the social and economic utility of the phenomenon versus harmful effects of DSN. There also exists a dilemma and lack of strategy in organizational use of DSN for business functions. There is a need for theoretical research to understand the underpinnings of DSN adoption and use followed by empirical investigation to explore the nuances of the theories. The social capital perspective of DSN discussed in this chapter offers a methodical approach to study DSN use.

REFERENCES

- Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *Academy of Management Review*, 27(1), 17–40. doi:10.5465/amr.2002.5922314
- Bapna, R., Gupta, A., Rice, S., & Sundararajan, A. (2017). Trust And The Strength Of Ties In Online Social Networks: An Exploratory Field Experiment. *Management Information Systems Quarterly*, 41(1), 115–130. doi:10.25300/MISQ/2017/41.1.06
- Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy*, 70(5), 9–49. doi:10.1086/258724
- Borgatti, S. P., & Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. *Journal of Management*, 29(6), 991–1013. doi:10.1016/S0149-2063(03)00087-4
- boyd D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.
- Burt, R. S. (1992a). *Structural Holes*. Harvard University Press.
- Burt, R. S. (1992b). The social structure of competition. In N. Nohria & R. Eccles (Eds.), *Networks and organizations: Structure, form, and action* (pp. 57–91). Harvard Business School Press.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94, S95–S120. doi:10.1086/228943

Dhar, S. (2018). *Role of Social Networking Sites in Creating Social Capital: Implications for Businesses* (Doctoral Dissertation). Aligarh Muslim University, Aligarh, India.

Dunbar, R. I. M., & Spoors, M. (1995). Social Networks, Support Cliques and Kinship. *Human Nature* (Hawthorne, N.Y.), 6(3), 273–290. doi:10.1007/BF02734142 PMID:24203093

Ellison, N. B., Gibbs, J. L., & Weber, M. W. (2015). The Use of Enterprise Social Network Sites for Knowledge Sharing in Distributed Organizations. *The American Behavioral Scientist*, 59(1), 103–123. doi:10.1177/0002764214540510

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook “friends:” Social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4), 1143–1168. doi:10.1111/j.1083-6101.2007.00367.x

Ellison, N. B., Steinfield, C., & Lampe, C. (2011). Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media & Society*, 13(6), 873–892. doi:10.1177/1461444810385389

Fu, F., Liu, L., & Wang, L. (2008). Empirical analysis of online social networks in the age of Web 2.0. *Physica A*, 387(2-3), 675–684. doi:10.1016/j.physa.2007.10.006

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380. doi:10.1086/225469

Hampton, K., & Wellman, B. (2003). Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City & Community*, 2(4), 277–311. doi:10.1046/j.1535-6841.2003.00057.x

Hanifan, L. J. (1916). The rural school community center. *The Annals of the American Academy of Political and Social Science*, 67(1), 130–138. doi:10.1177/000271621606700118

Helliwell, J. F., & Putnam, R. D. (2004). The social context of well-being. *Philosophical Transactions of the Royal Society*, 359(1449), 1435–1446. doi:10.1098/rstb.2004.1522 PMID:15347534

Hofer, M., & Aubert, V. (2013). Perceived bridging and bonding social capital on Twitter: Differentiating between followers and followees. *Computers in Human Behavior*, 29(6), 2134–2142. doi:10.1016/j.chb.2013.04.038

Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. P. (2014). What’s Different about Social Media Networks? A Framework and Research Agenda. *Management Information Systems Quarterly*, 38(1), 274–304. doi:10.25300/MISQ/2014/38.1.13

Kane, G. C., & Borgatti, S. P. (2011). Centrality–Is Proficiency Alignment And Workgroup Performance. *Management Information Systems Quarterly*, 35(4), 1063–1078. doi:10.2307/41409973

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68. doi:10.1016/j.bushor.2009.09.003

Kautz, H., Selman, B., & Shah, M. (1997). ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3), 63–65. doi:10.1145/245108.245123

Koka, B. R., & Prescott, J. E. (2001). Strategic alliances as social capital: A multidimensional view. *Strategic Management Journal*, 23(9), 795–816. doi:10.1002/mj.252

Lee, J., & Suh, E. (2013, June). *An Empirical Study of the Factors Influencing Use of Social Network Service*. PACIS, Jeju Island, Korea.

Liben-Nowell, D., & Kleinberg, J. (2007). The Link-Prediction Problem for Social Networks. *Journal of the American Society*, 58(7), 1019–1031.

Lipsman, A., Mud, G., Rich, M., & Bruich, S. (2012). Beyond the “Like” Button: The Impact of Mere Virtual Presence on Brand Evaluations and Purchase Intentions in Social Media Settings. *Journal of Advertising Research*, 52(1), 40–52. doi:10.2501/JAR-52-1-040-052

Moreno, J. L. (1934). *Who Shall Survive?* Nervous and Mental Disease Publishing Company.

Morris, M. R., Teevan, J., & Panovich, K. (2010, April). *What Do People Ask Their Social Networks, and Why*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, New York, NY. 10.1145/1753326.1753587

Pool, I. D. S., & Kochen, M. (1979). Contacts and Influence. *Social Networks*, 1(1), 5–51. doi:10.1016/0378-8733(78)90011-4

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. Simon & Schuster Paperbacks.

Putnam, R. D. (2007). E Pluribus Unum: Diversity and Community in the Twenty-first Century The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, 30(2), 137–174. doi:10.1111/j.1467-9477.2007.00176.x

Rowley, T. J. (1997). Moving beyond dyadic ties: A network theory of stakeholder influences. *Academy of Management Review*, 22(4), 887–910. doi:10.5465/amr.1997.9711022107

Seibert, S. E., Kraimer, M. L., & Liden, R. E. (2001). A Social Capital Theory of Career Success. *Academy of Management Journal*, 44(2), 219–237.

Tsai, W. (2000). Social Capital, Strategic Relatedness and the Formation of Intraorganizational Linkages. *Strategic Management Journal*, 21(9), 925–939. doi:10.1002/1097-0266(200009)21:9<925::AID-SMJ129>3.0.CO;2-I

Uslaner, E. M. (2000). Social Capital and the Net. *Communications of the ACM*, 43(12), 60–64. doi:10.1145/355112.355125

Wasko, M. M. L., & Faraj, S. (2005). Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *Management Information Systems Quarterly*, 29(1), 35–57. doi:10.2307/25148667

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, 22(1), 213–238. doi:10.1146/annurev.soc.22.1.213

Williams, D. (2006). On and Off the ‘Net: Scales for Social Capital in an Online Era. *Journal of Computer-Mediated Communication*, 11(2), 593–628. doi:10.1111/j.1083-6101.2006.00029.x

Woolcock, M. (1998). Social capital and economic development: Toward a theoretical synthesis and policy framework. *Theory and Society*, 27(2), 151–208. doi:10.1023/A:1006884930135

ADDITIONAL READING

Burke, M., Kraut, R., & Marlow, C. (2011, May). *Social capital on Facebook: Differentiating uses and users*. Paper presented at the ACM CHI 2011. New York. 10.1145/1978942.1979023

Burt, R. S. (2000). The network structure of social capital. *Research in Organizational Behavior*, 22, 345–423. doi:10.1016/S0191-3085(00)22009-1

Chen, R., & Sharma, S. K. (2013). Self-disclosure at social networking sites: An exploration through relational capitals. *Information Systems Frontiers*, 15(2), 269–278. doi:10.1007/10796-011-9335-8

Ellison, N., Vitak, J., Gray, R., & Lampe, C. (2014). Cultivating Social Resources on Social Network Sites: Facebook Relationship Maintenance Behaviors and Their Role in Social Capital Processes. *Journal of Computer-Mediated Communication*, 19(4), 855–870. doi:10.1111/jcc4.12078

Hanneman, R. A., & Riddle, M. (2005). *Introduction to Social Network Methods*. University of California, Riverside. Published in digital form at <http://faculty.ucr.edu/~hanneman/>

KEY TERMS AND DEFINITIONS

Digital Social Network: Digital platforms that facilitate social networking, where individuals or groups (nodes) can articulate their profile and connections discoverable in the network and digitally interact with other nodes in the network.

Social Capital: Denotes goodwill, fellowship, help and sympathy among connections in a social network that yields benefits such as support, solidarity, and access to resource.

Social Network: Social relations studied as a network, where individuals or groups are represented as nodes and their relationship and interactions are represented by edges.

Digital Documents

Anastasios Dimou

Independent Researcher, Athens, Greece

Apostolos Syropoulos

 <https://orcid.org/0000-0002-9625-1482>

Greek Molecular Computing Group, Greece

INTRODUCTION

Nowadays documents (e.g., books, articles, papers, etc.) are created electronically and they are printed electronically. Modern printers are programmable and so documents are uploaded to them as *programs* in some *page description language* (like PostScript, PDF, and PCL). Currently, PDF is the de facto industrial standard for printable documents so all documents are finally converted to PDF.

Definition A digital document is one that is created exclusively with digital means (e.g., computer programs), can be read with digital means (e.g., electronic book readers), and can be transferred to paper by purely digital means (e.g., fonts are digital and the printing machine transfers the digital document to paper).

Documents are created using either sophisticated GUIs or using an “advanced” text editor. Programs like MS Word and LibreOffice Writer are sophisticated GUIs that are used to create digital documents. On the other hand, one can create the same documents “manually” by using a simple text editor and/or some CLI tools. For example, it is entirely possible to create a document in the Open Document Format for Office Applications (ODF) with such tools and of course most people create their LaTeX documents using simple or advanced text editors like NotePad++.

The various file formats that are used to electronically encode documents based on the idea of *markup*. To understand what markup is consider a book that submitted for publication. Typically, a copy editor will annotate the text with various types of instructions. This is markup. In case, the markup is entered electronically, then, clearly, it is called *electronic* markup. *Specific* markup is a form of markup that tells the formatted to do each point. *Generic* markup adds information about the logical components of a document. In order to add markup to a text we need a precisely defined notation. Such a notation is called a *markup language*. The basic elements of such a language are called tags.

A markup language defines *tags* that group text and affect the way specific parts of a document will be rendered. In certain cases it is possible to define *macros*, that is, new tags that have the combined functionality of *primitive* tags. There are several markup languages that are widely used today. These include TeX, DocBook, and HTML and macro packages like LaTeX. However, in a way all these markup languages are descendants of SGML.

In what follows we will give a brief overview of SGML and XML. Then, we will discuss page description languages in some details, and we will continue with an overview of some very popular markup languages.

DOI: 10.4018/978-1-7998-3479-3.ch077

BACKGROUND

The *Standard Generalized Markup Language* (SGML) is the International Organization for Standardization (ISO) standard for document description [see (ISO-Standard, 1986) and (van Herwijnen, 1990)]. SGML is actually a meta-language that allows people to create markup languages. Typically, an SGML document consists of a file that contains marked-up data, the SGML declaration, and the document type definition (DTD). The SGML declaration specifies which characters and delimiters are used. For example, it is common to use the characters <, >, and /> as delimiters. And this why it is so common to see tags like <intro> and </intro>. Also, the The SGML declaration specifies which character set (Unicode or other) is used. Figure 2 shows an SGML Declaration for HTML 3.2. The DTD specifies how one can markup a class of documents. In particular, it defines the structure of a document and it is written in SGML. In general, no two different document classes have the same structure. Figure 1 shows a very simple DTD.

Figure 1. A very simple DTD from (Syropoulos, Tsolomitis, & Sofroniou, 2003).

```
<!ELEMENT first_name (#PCDATA)>
<!ELEMENT last_name (#PCDATA)>
<!ELEMENT profession (#PCDATA)>
<!ELEMENT name
(first_name, last_name)>
<!ELEMENT person
(name, profession*)>
<!ELEMENT scientists (person*)>
```

Today there are a few projects that are directly based on SGML. The *Text Encoding Initiative* (TEI) is such a project whose aim was to create software-independent methods for encoding humanities data in electronic form (TEI: Text Encoding Initiative). Docbook is another markup language based on SGML that was designed for technical documentation (DocBook Version 5.2, 2019). However, the current version of DocBook is an XML application.

The eXtensible Markup Language (XML) [see (Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 2008) and (Harold & Means, 2009)] is a subset of SGML. The goal of XML is “to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML.” XML has been used to design the file formats of many office-productivity tools including the Office Open XML format (Standard ECMA-376, 2016) and the Open Document Format for Office Applications (OASIS Standards, 2011). The Open Document Format is also the ISO/IEC 26300:2006 standard.

An XML document consists of a DTD and a data file. For example, if one uses a DTD like the one shown in figure 1, then she can create a valid XML document like the one shown in figure 3. Note that the various <something> and </something> pairs are called *elements*. One of the differences between SGML and XML is that the later uses Unicode’s UTF-8 encoding as the default character encoding. In addition, since XML is a subset of SGML, it does not “understand” a number of SGML declarations, constructs, and entity declarations.

Figure 2. An SGML Declaration for HTML 3.2 available from (Sample SGML Open Catalog for HTML 3.2).

<pre> <!SGML "ISO 8879:1986" CHARSET BASESET "ISO 646:1983//CHARSET International Reference Version (IRV)//ESC 2/5 4/0" DESCSET 0 9 UNUSED 9 2 9 11 2 UNUSED 13 1 13 14 18 UNUSED 32 95 32 127 1 UNUSED BASESET "ISO Registration Number 100//CHARSET ECMA-94 Right Part of Latin Alphabet Nr. 1//ESC 2/13 4/1" DESCSET 128 32 UNUSED 160 96 32 CAPACITY SGMLREF TOTALCAP 200000 GRPCAP 150000 ENTCAP 150000 SCOPE DOCUMENT SYNTAX SHUNCHAR CONTROLS 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 127 BASESET "ISO 646:1983//CHARSET International Reference Version (IRV)//ESC 2/5 4/0" DESCSET 0 128 0 FUNCTION RE 13 RS 10 SPACE 32 </pre>	<pre> TAB SEPCHAR 9 NAMING LCNMSTRT "" UCNMSTRT "" LCNMCHAR ".-" UCNMCHAR ".-" NAMECASE GENERAL YES ENTITY NO DELIM GENERAL SGMLREF SHORTREF SGMLREF NAMES SGMLREF QUANTITY SGMLREF ATTSPLEN 65536 LITLEN 65536 NAMELEN 65536 PILEN 65536 TAGLVL 100 TAGLEN 65536 GRPGTCNT 150 GRPCNT 64 FEATURES MINIMIZE DATATAG NO OMITTAG YES RANK NO SHORTTAG YES LINK SIMPLE NO IMPLICIT NO EXPLICIT NO OTHER CONCUR NO SUBDOC NO FORMAL YES APPINFO NONE> </pre>
--	---

Figure 3 A simple XML document that is based on the DTD shown in figure 1.

```

<?xml version="1.0" standalone="no"?>
<!DOCTYPE scientists SYSTEM
"http://ocean1.ee.duth.gr/dtds/scientists.dtd">
<scientists>
  <person>
    <name>
      <first_name>Alan</first_name>
      <last_name>Turing</last_name>
    </name>
    <profession>Computer Scientist</profession>
    <profession>Mathematician</profession>
    <profession>Cryptographer</profession>
  </person>
  <person>
    <name>
      <first_name>Leslie</first_name>
      <last_name>Lampport</last_name>
    </name>
    <profession>Computer Scientist</profession>
  </person>
</scientists>

```

Figure 4 Sample PostScript file.

```

%!PS
/Symbol findfont
85 scalefont setfont
10 500 moveto
(APOSTOLOS) show
10 400 moveto
(SUROPOULOS) show
showpage

```

Figure 5. A sample PDF file that was automatically created from the file shown in figure 4.

<pre> %PDF-1.7 %çóÍÖ 1 0 obj <</Type /Catalog /Extensions <</ADBE <</ExtensionLevel 8 /BaseVersion /1.7 >> >> /Pages 2 0 R /Metadata 3 0 R >> endobj 4 0 obj <</CreationDate (D:20190526192107+03'00') /ModDate (D:20190526210559+03'00') /Producer (GPL Ghostscript 9.19) >> endobj 3 0 obj <</Subtype /XML /Type /Metadata /Length 3100 >>stream <?xml:space="preserve" id="W5M0MPC"?> ... Many lines of XML Data deleted ... <?xml:space="preserve"?>endstream endobj 2 0 obj <</Count 1 /Type /Pages /Kids [5 0 R] >> endobj 5 0 obj <</Rotate 0 /Parent 2 0 R /Contents 6 0 R /Type /Page </pre>	<pre> /MediaBox [0 0 612 792] /CropBox [0 0 612 792] /Resources <</ProcSet [/PDF /Text] /Font <</T1_0 7 0 R >> >> endobj 7 0 obj <</Encoding 8 0 R /Subtype /Type1 /Type /Font /BaseFont /Symbol >> endobj 8 0 obj <</BaseEncoding /WinAnsiEncoding /Differences [65 /Alpha 76 /Lambda 79 /Omicron /Pi 82 /Rho /Sigma /Tau /Upsilon </pre>	<pre>] /Type /Encoding >> endobj 6 0 obj <</Length 9 0 R >>stream BT /T1_0 85 Tf 10 500 Td (APOSTOLOS)Tj 0 -100 TD (SUROPOULOS)Tj ET endstream endobj 9 0 obj 67 endobj xref 0 10 0000000000 65535 f 0000000015 00000 n 0000003450 00000 n 0000000271 00000 n 0000000143 00000 n 0000003508 00000 n 0000003926 00000 n 0000003694 00000 n 0000003776 00000 n 0000004044 00000 n trailer <</Root 1 0 R /ID [(\270p\326G\237ysM\233R\030\231\204\205\324!) (e\345W*\177\267\350K\265\233\007M\251\031Q)] /Size 10 /Info 4 0 R >> startxref 4062 %%EOF </pre>
---	---	---

PAGE DESCRIPTION LANGUAGES

A page description language (PDL) describes the appearance of a printed page not in the form of an image file (e.g., a PNG file) but in notation specifically designed for this purpose. In general, a PDL is used to describe what goes on each page of a document (letters, graphics, color, etc.). There are many PDLs but three of them are still used widely today: PostScript, PDF, and PCL.

PostScript

PostScript® (PostScript Language Reference, 3rd Edition, Adobe) is a simple interpretive programming language with powerful graphics capabilities. A program in this language is a description of document that is generated by composition system and which can be transferred to printing device or used to have a look at the final product. The description is high-level and device-independent. PostScript is Turing-complete, which means that it can be used to perform a wide range of computing tasks.

PDF The Adobe Portable Document Format (PDF) is the native file format of the Adobe® Acrobat® family of products. These products have been designed so to allow users to exchange and view electronic documents easily and reliably, independently of the environment in which they were created. In 2008, PDF-1.7 became an ISO standard (ISO 32000-1:2008, 2008). As PDF has become the lingua franca of the printing industry, nowadays, most printing devices can natively accept and process PDF documents. A PDF file is structured as a series of possible nested objects and streams. Figure 5 shows a PDF file that corresponds to the PostScript file shown in figure 4. Note that the PDF file has been transformed so make visible all relevant details and also it has been edited so to fit on a single page, nevertheless, the structure of the file is pretty obvious.

Although the PDF format is extremely popular today, still we cannot be sure whether someone in 100 years from now will be able to read a PDF file! This is the reason ISO created a few standards that describe a version of PDF specialized for use in the archiving and long-term preservation of electronic documents. These standards are collectively known as PDF/A.

PCL

PCL is a PDL created by Hewlett-Packard (PCL5 Printer Language Technical Reference Manual: Part I, 1992). The language was designed so to provide an economical and efficient way for application programs to control a range of printer features across a number of printing devices. PCL commands are compact escape sequence codes that are embedded in the data stream that is supplied to the printing device. This simply means that PCL files are non-human readable in any possible way. Having a binary format minimizes both data transmission and command decoding overhead. Although this feature was particularly appealing in the 1990s, today it is almost irrelevant. And this is why PCL has remained a PDF that has not gained the status of PostScript and PDF.

XML-BASED DOCUMENT FORMATS

As was noted in the background section, there are many XML-based document formats. In particular, both Microsoft Office and LibreOffice, which are the most popular office productivity suites, use as their native file format XML-based files. In figure 6 we show the simplest possible workbook that can be opened by Excel. However, the native format that these productivity suits accept is a bit more complex than this simple file. In particular, the native document files produced by Microsoft Office have the .docx file extension, while native document file produced by LibreOffice have the .odt file extension.

Figure 6. The simplest possible workbook that can be opened by Excel.

```
<?xml version="1.0"?>
<?mso-application progid="Excel.Sheet"?>
<Workbook
  xmlns="urn:schemas-microsoft-com:office:spreadsheet"
  xmlns:ss="urn:schemas-microsoft-com:office:spreadsheet">
  <Worksheet ss:Name="Sheet1">
  </Worksheet>
</Workbook>
```


Both file types are in fact archives or, more specifically, zipped files. Figure 7 shows the files and folders that are used to save all required information. Naturally, if a document includes images or any other permissible media file, these are stored in the archive.

Figure 7. Directory structure of common XML-based document files.

Microsoft Word .docx	LibreOffice Writer .odt
[Content_Types].xml	mimetype
_rels/.rels	Thumbnails/thumbnail.png
word/_rels/document.xml.rels	content.xml
word/document.xml	styles.xml
word/theme/theme1.xml	settings.xml
word/settings.xml	meta.xml
word/fontTable.xml	manifest.rdf
word/webSettings.xml	Configurations2/toolpanel/
docProps/app.xml	Configurations2/accelerator/
docProps/core.xml	Configurations2/floater/
word/styles.xml	Configurations2/statusbar/
	Configurations2/progressbar/
	Configurations2/toolbar/
	Configurations2/images/Bitmaps/
	Configurations2/popupmenu/
	Configurations2/menubar/
	META-INF/manifest.xml

DocBook has been designed to allow people prepare technical documents easily. The current version of DocBook (version 5.2) is an XML application. Unlike the office productivity file formats, it is relatively easy to manually create a DocBook document. The first thing one has to specify in a plain text file that will contain the markup is the file header. The header for version 5.0 of DocBook is shown below:

DocBook elements can be divided broadly into these categories:

- A *set* may contain one or more books.
- A *book* consists of a dedication page, a table of contents, parts, chapters, etc. Chapters are considered *components* while parts and reference sections are considered *divisions*.
- *Sections*, which subdivide components. There are several sectioning elements: from <sect1> to <sect5>. In addition, one can use the <section> element which can be nested, that is, one can have a <section> into another <section>, etc.
- *Meta-information* elements that provide the <title> of a document, its <author>s, etc., and are included in an <info> element.
- *Block* elements appear immediately below component and sectioning elements. They can be divided into a number of categories: lists, admonitions, line-specific environments, synopses of several sorts, <table>s, <figure>s, <example>s, etc.
- *Inline* elements are used to mark up running text. For example, they can be used to change the use of a font, but they do not cause line or paragraph breaks.

Figure 9 shows the skeleton of a complete DocBook article.

Figure 8. The header of a typical DocBook file

```
<?xml version='1.0'?>
<!DOCTYPE book PUBLIC "-//OASIS//DTD DocBook V5.0//EN"
"http://www.oasis-open.org/docbook/xml/5.0/docbook.dtd">
```

There are several methods to print a DocBook document. One of them makes use of eXtensible Stylesheet Language Formatting Objects (XSL-FO) (Pawson, 2002). The user invokes `xsltproc`, a command line XSLT processor. The XLST (eXtensible Stylesheet Language Transformations) is a language for transforming XML documents into other XML documents or other formats (e.g., HTML or XSL-FO). The final XSL-FO file can be printed using an XML-FO processor [see (XSL-FO Processors, 2014) for a list of XML-FO processors].

XHTML is a reformulation of HTML 4 as an XML 1.0 application (XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition), 2002). HTML 4 is an SGML application as were all previous versions of HTML. Since the current version of HTML is HTML 5, we will not say more about either XHTML or previous versions of HTML.

WEB-ORIENTED DOCUMENT FORMATS

There are a few more notable document formats that are used widely for the creation of digital documents.

HyperText Markup Language 5

The prefix hyper- generally means “above” or “beyond,” thus hypertext is something that goes beyond the limitations of ordinary text (e.g., it can be non-sequential). Hypertext is text which contains links to other information. The term was coined by Theodor Holm “Ted” Nelson around 1963. Software programs such as dictionaries and encyclopedias have long used hypertext in their definitions allowing readers to quickly navigate content. HyperMedia is a term used for hypertext which is not constrained to be text: it can include graphics, video and sound, etc. Ted Nelson coined this term too.

Tim Berners-Lee created what we call the Web while he was working at CERN. He created HTML and the HyperText Transfer Protocol (HTTP) and designed and implemented the first web browser. On October 1991 “HTML Tags” was made available. A description of version 2.0 of HTML was published on November 1995. HTML 3.2 was published on January 1997. HTML 4.0 was published on December 1997. HTML5 was published on October 2014 while HTML 5.1 was published on November 2016. The latest version of HTML5 is HTML 5.3 and its specification was published on 18 October 2018 (HTML 5.3: W3C Recommendation, 2018). Figure 10 shows a sample HTML5 file. Let us briefly describe the structure of an HTML5 file.

An HTML5 file consists of three parts: the `<header>`, the `<body>`, and a document identification line. The last part is just `<!DOCTYPE HTML>`. The `<header>` is where authors put the title of the document, meta-information (e.g., which character set is the default, keywords, etc.), and where the document loads external files that provide additional functionality [e.g., JavaScript code or Cascading Style Sheets, which is a “simple mechanism for adding style (e.g., fonts, colors, spacing) to Web documents” (Cascading Style Sheets: Home Page, 2019)]. The `<body>` of the document contains the content of the

Figure 9. The skeleton of a DocBook file

```

<?xml version='1.0'?>
<!DOCTYPE book PUBLIC "-//OASIS//DTD DocBook V5.0//EN"
    "http://www.oasis-open.org/docbook/xml/5.0/docbook.dtd">

<article>
  <info>
    <title>My Article</title>
    <author><honorific>Dr</honorific><firstname>Apostolos</firstname>
      <surname>Syropoulos</surname></author>
  </info>
  <para>... </para>
  <section><title>On the Possibility of Going Home</title>
    <para>... </para>
  </section>
  <bibliography>... </bibliography>
</article>

```

document. Note that we say the content and not just the text since one can have just video, an audio clip, an animation, a computer game, etc., in the <body> of an HTML document.

The first versions of HTML were designed so to be able to render text with some pictures. However, it was soon made clear that one could have videos, animations, sounds, etc., in a Web page. Since it was not possible to directly add this content, many software developers and companies created browser extensions or plugins that allowed different kinds of content to be included in a Web page. Unfortunately, not all plugins worked on all computer platforms. For example, the Adobe Flash plugin is not available for Android thus making it impossible for millions of users to browse Web pages that include Flash content. This and other similar reasons was the driving force behind HTML5, which includes all relevant tools and technologies to create Web pages that do not depend on external plugins.

In general, HTML documents can be interactive since one can embed JavaScript code thus “allowing” a Web page to react to user input (i.e., mouse movements, keyboard strokes, etc.). This of course is totally new form of document and one that is enhanced every single day.

ePUB (EPUB 3.2, 2019) is a document format similar to Microsoft’s .docx. The difference is that an ePUB archive consists of HTML files. One can say that an ePUB file is actually a Web site in a box. There are several tools to create ePUBs including Google Docs and the tex4ebook tool-set.

Figure 10. Sample HTML5 file. The arrows are used to annotate each line

```

<!DOCTYPE html> ← document type
<html> ← beginning of document
<head> ← beginning of head
<meta charset="UTF-8">
<title>Greetings</title> ← the title
</head> ← end of head
<body> ← beginning of body
Hello World! ← document contents
</body> ← end of body
</html> ← end of document

```

LaTeX

LaTeX [see (Lamport, 1994) and (Syropoulos, Tsolomitis, & Sofroniou, 2003)] is document preparation system that runs on top of TeX (Knuth, 1984), which is a typesetting language designed and implemented by Donald E. Knuth. Document preparation using LaTeX is task where the user concentrates on the logical structure of the document and not on how each part should look like. The latest approach is known in computer jargon as WYSIWYG (what you see is what you get). Clearly, the logical approach to document preparation is employed when preparing documents using DocBook, HTML, TEI, etc. Figure 10 shows a typical XeLaTeX source file. This is a modern variant of LaTeX where one has to use XeTeX to process input files. This is a modern variant of TeX that can natively accept Unicode input and that can use OpenType fonts, something not possible until some years ago. When preparing a LaTeX document it is advisable to use some special editor that can assist you in the preparation process. However, there are a few GUI systems that allow one to prepare a document and which emit LaTeX files. Such a system is LyX (LyX: The Document Processor, 2018) and figure 12 shows how the LyX's GUI looks like.

Figure 11. A typical XeLaTeX input file

```
\documentclass[a4paper]{article}
\usepackage{xltextra}
\usepackage{unicode-math}
\begin{document}
\setmainfont[Mapping=tex-text,Ligatures=Common]{Minion Pro}
\setsansfont[Mapping=tex-text,Scale=MatchLowercase]{Myriad Pro}
\setmonofont[Scale=MatchLowercase]{UM Typewriter}
\setmathfont[Scale=MatchUppercase]{Asana Math}
\title{My Title}
\author{Anastasios Dimou}
\maketitle
\secgction{A section}
\begin{displaymath}
\alpha^2+\beta^2=\gamma^2.
\end{displaymath}
\end{document}
```

Similar to an HTML document, a LaTeX document has three parts (see figure 11). The first part is the header where one declares the document type she intends to prepare, the default font size, paper size, etc. For example, the following header

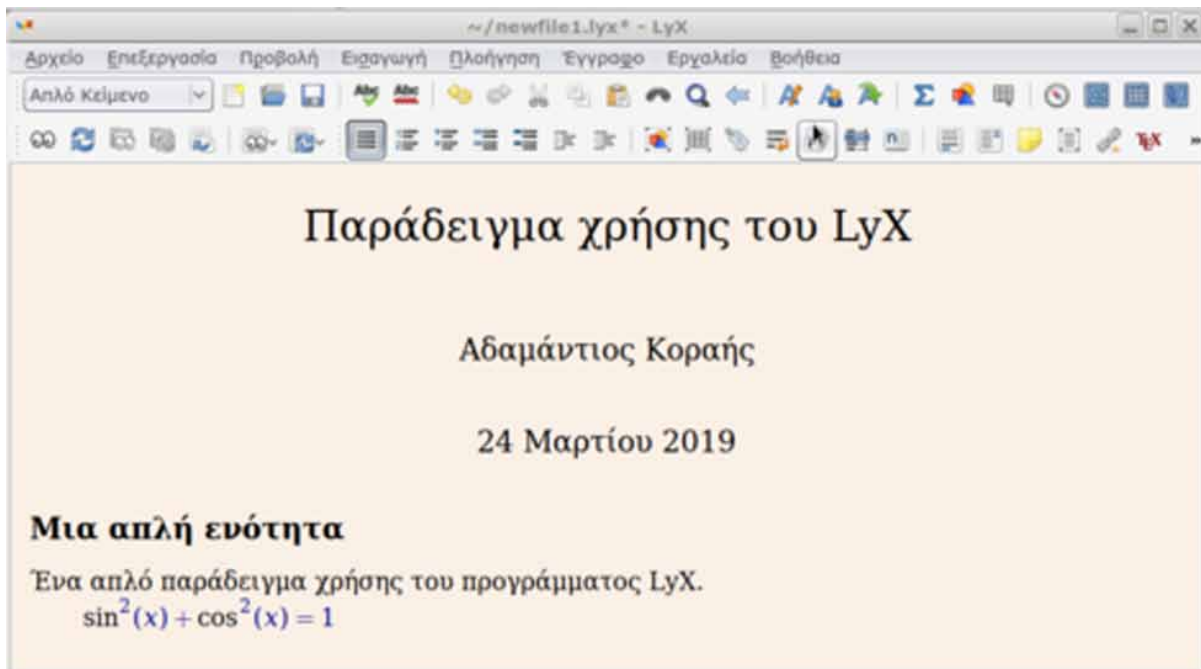
```
\documentclass[a4paper,11pt]{article}
```

specifies that we are going to prepare an article on a A4 paper using 11 point as the default font size. What goes between the header and the command that follows

```
\begin{document}
```

is called preamble. There the user specifies which external macro packages should be loaded. These packages provide additional functionality to LaTeX and can be used to specify a custom paper and text

Figure 12. A screenshot of LyX in action



size, to modify the final layout, etc. Finally, what goes after the `\begin{document}` command is our text interspersed with formatting commands and/or structural markup commands.

As can be seen in figure 10, LaTeX markup is bit different but the principle is the same. Wikis are also written in a rather “strange” markup language.

Wikis

Wikis are Web pages where people write about something and others are usually allowed to modify the content in order to improve its quality. The idea was pioneered by Wikipedia, the free encyclopedia (wiki is Hawaiian for “fast; quick”). Wiki markup (Help:Wikitext, 2019) is a simplified form of markup. For example, heading are specified as follows:

```
==== Heading 4 ====
```

What goes inside the equal signs is the heading and the number of equal sign should be used to write sections, subsections, etc. A heading with only one equal sign, should be used for the title of a document. To enter a horizontal rule, one has to enter `----`. In HTML one get the same by entering `<hr/>`. In addition, one can use a number of HTML elements to format a Wiki page.

SOLUTIONS AND RECOMMENDATIONS

There are many different electronic documents types that have been designed to ease the process of creating such documents. Opting to use one tool or one system over some other is purely a matter of taste.

Some people feel more comfortable with GUIs whereas others prefer to write documents and mark up together. However, for long documents and very serious work, we believe the second approach is more reasonable. After all, one should care about she writes and how she writes something.

FUTURE RESEARCH DIRECTIONS

Many problems related to the preparation of digital documents have been properly solved. However, there are still problems that have not been solved completely or properly. For example, the preparation of traditional Mongolian text is something that needs our attention. Also, it is quite possible that new types of media may appear and these should find their way in our electronic documents. The advent of JSON may make it necessary to create document formats based on this technology. *Active documents* (Kohlhase, 2012), that is, digital documents augmented with embedded services that activate the content for interaction and adaption, is another interesting research direction. These documents reify the knowledge contained in a document. Thus a mathematical active document could include exercises that could be solved by the reader while reading the document. In addition, the document reader could assist the person reading the document to solve the exercises.

CONCLUSION

We have presented digital documents and their general characteristics. Also, we described in some detail the most important formats of digital documents and how one can structure such a document. Naturally, we omitted to mention a few formats (e.g., the Rich Text Format), however, we concentrated on what we thought are the most important milestones.

REFERENCES

- Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. Retrieved 05 26, 2019, from <https://www.w3.org/TR/xml/>
- Cascading Style Sheets. (2019). Retrieved 05 28, 2019, from Cascading Style Sheets: Home Page: <https://www.w3.org/Style/CSS/>
- DocBook Version 5.2. (2019). Retrieved 05 27, 2019, from Oasis: <https://tdg.docbook.org/tdg/5.2/>
- EPUB 3.2. (2019). Retrieved 05 29, 2019, from <https://w3c.github.io/publ-epub-revision/epub32/spec/epub-spec.html>
- Harold, E. R., & Means, W. S. (2009). *XML in a Nutshell: A Desktop Quick Reference*. O'Reilly Media.
- Help:Wikitext. (2019). Retrieved 05 29, 2019, from Wikipedia: The free encyclopedia that anyone can edit: <https://en.wikipedia.org/wiki/Help:Wikitext>
- HTML 5.3: W3C Recommendation. (2018). Retrieved 05 29, 2019, from <https://www.w3.org/TR/html53/>
- ISO 32000-1:2008. (2008). Retrieved 05 26, 2019, from International Organization for Standardization: <https://www.iso.org/standard/51502.html>

ISO-Standard. (1986). *Technical Committee: ISO/IEC JTC 1/SC 34 Document description and processing languages*. Retrieved 05 25, 2019, from <https://www.iso.org/standard/16387.html>

Knuth, D. E. (1984). *The TeXbook*. Addison-Wesley Professional.

Kohlhase, M. (2012). The Flexiformalist Manifesto. In *2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (pp. 30-35). Timisoara, Romania: IEEE.

Lamport, L. (1994). *LaTeX: A Document Preparation System*. Addison-Wesley Professional.

LyX: The Document Processor. (2018). Retrieved 05 29, 2019, from LyX: The Document Processor: <https://www.lyx.org/>

OASIS Standards. (2011). Retrieved 05 26, 2019, from OASIS. Open standards. Open source.: <https://www.oasis-open.org/standards#opendocumentv1.2>

Pawson, D. (2002). *XSL-FO: Making XML Look Good in Print*. O'Reilly Media.

PCL5 Printer Language Technical Reference Manual: Part I. (1992). Retrieved 05 26, 2019, from <http://www.hp.com/ctg/Manual/bpl13210.pdf>

PostScript Language Reference, 3rd Edition. (n.d.). Retrieved 05 26, 2019, from A: <https://www.adobe.com/content/dam/acom/en/devnet/actionsript/articles/PLRM.pdf>

Sample SGML Open Catalog for HTML 3.2. (n.d.). Retrieved 05 26, 2019, from <https://www.redfox-center.org/HTML/sgml4html32.htm>

Standard ECMA-376. (2016, 12 4). Retrieved 05 26, 2019, from ECMA International: <http://www.ecma-international.org/publications/standards/Ecma-376.htm>

Syropoulos, A., Tsolomitis, A., & Sofroniou, N. (2003). *Digital Typography Using LaTeX*. Springer-Verlag New York.

TEI: Text Encoding Initiative. (n.d.). Retrieved 05 26, 2019, from <https://tei-c.org/>

van Herwijnen, E. (1990). *Practical SGML*. New York: Springer US.

XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition). (2002). Retrieved 05 27, 2019, from <https://www.w3.org/TR/2002/REC-xhtml1-20020801/>

XSL-FO Processors. (2014). Retrieved 05 27, 2019, from W3C Wiki Page: https://www.w3.org/community/ppl/wiki/XSL-FO_Processors

KEY TERMS AND DEFINITIONS

CLI: Command line interface is a way to run programs from a Unix shell or the Windows command prompt.

GUI: Graphical user interface is a graphical interface that is used to communicate and operate a computer program. Most computer programs come with a GUI since almost all users are accustomed to the use of the mouse.

Markup: A technology that has been developed for creation of digital documents. The text of a digital document is interspersed with markup elements that are either semantic or typographic annotations.

Digital Media and New Forms of Journalism

Lambrini Papadopoulos

 <https://orcid.org/0000-0002-0968-8046>

Aristotle University of Thessaloniki, Greece

Theodora A. Maniou

 <https://orcid.org/0000-0003-4299-5221>

University of Cyprus, Cyprus

INTRODUCTION

This chapter analyses the role and operation of digital media, the impact of digital technology on the media sector and new forms of journalism that have emerged in the digital era. With its origins in the 1990s, digital technology has been embedded in the media sector for over two decades. These crucial years saw the rapid development and evolution of *digital technology* to *smart technology* and then to the *interactive format* of Web 2.0, which decisively altered media operations, content and format. This chapter aims to define and describe *digital media* and other related terms, while also shedding light on *new forms of journalism* that arise from the widespread adoption of digital technology. These may include (but are not limited to) *Virtual Reality (VR) journalism*, *drone journalism*, *selfie journalism* and *robot journalism*. Finally, this chapter offers an academic overview and understanding of the issues surrounding the way in which technological disruption transforms old habits and practices in the newsrooms, leading to innovative storytelling that transcends time and space.

BACKGROUND

The term *digital media* is used to refer to digitalized (information or other) content that can be transmitted via the internet and/or a computer network. Initially, this content could (potentially) include text, audio, video and/or graphics, although since the early 2000s this variety of elements has been enriched by new forms of journalism emerging from a range of technological trends, particularly the rapid spread of *smart technology*.

In the early days of Web 2.0, users could locate media content through the internet; however, this content offered the same information, in terms both of quantity and quality, as the traditional media, whereas new content was available only via paid subscription platforms (van der Wurff, 2008). Since then, media have converged at spectacular speed: from smartphones to radios, television sets to tablets, newspapers to computers, the audience increasingly moves between an ever-extending menu of media platforms (Cushion & Sambrook, 2016). The notion of digitally converged media technologies has been the object of widespread attention since the 1990s, initially focused on the convergence of broadcast television and the networked computer, later evolving to include all media entities.

One of the most widely observed consequences of the growth in digital media is audience fragmentation. As more offerings are delivered on broadband networks and more choices are available on-demand,

DOI: 10.4018/978-1-7998-3479-3.ch078

patterns of consumption become more widely distributed (Webster & Ksiazek, 2012). Furthermore, media convergence in the digital age is frequently used both in the academic field and the media industry to denote the ongoing restructuring of media companies as well as to describe developments in media forms, distribution and consumption. The use of the concept has developed from being mainly connected with digitalization in media technology to also include elements of the ongoing process of integration, combination, competition and divergence (Appelgren, 2004).

Although the rapid evolution of digital media raised questions regarding the future survival of traditional media, the latter still constitute a significant part of the media system. The co-existence of traditional and digital media brought forward the notion of *hybridity*. Hybrid media, developed in the late 1990s, due to the elusive progress of digital technology, rely heavily on interaction and could be described as the result of the deep remixing of previously separate media techniques and languages (Manovich, 2007). Chadwick, Dennis & Smith (2016) describe the hybrid media system as a system built on interactions among older and newer media logics, where logics are defined as bundles of technologies, genres, norms, behaviors and organizational forms. Actors in this system are articulated by complex and ever-evolving relationships based on adaptation, interdependence, concentrations and diffusions of power. Actors create, tap or steer information flows in ways that suit their goals and in ways that modify, enable or disable others' agency, across and between a range of older and newer media settings (Chadwick, 2013, p.4).

FOCUS OF THE ARTICLE

Until recently, most digital media were focused on converting analog data/footage into digital. However, in recent years they have developed distinctive capabilities of content management and became more accessible to a growing global audience that seeks information online. According to the Digital Future Project annual study (2018), 49% percent of users go online to look for news at least daily, and 70% of users go online for news at least weekly. Moreover, the emergence of social media as consumers' primary news source, the extensive use of mobile platforms and the advent of complex technologies such as Virtual Reality (VR) and Artificial Intelligence (AI) are paving the way for new forms of journalism that are shaping the future of the industry. Apart from VR and AI, drones and bots are also expected to play an important role in the nearby future by exploring new reporting frontiers. Drones are offering unique visual perspectives and assist journalists in filtering and authenticating their news stories whereas bots are changing the way news are delivered by providing voice messages, news bulletin or even breaking news to consumers (Veglis & Maniou, 2019). To this end, digital media are evolving in a dynamic way, entering a new era, where new technological trends such as drones, bots and smart glasses bring new opportunities for experiencing and exploring new kinds of journalism with distinctive features, traits and values.

VR (Virtual Reality) Journalism

A relatively recent development in the realm of digital journalism is *VR(virtual reality) journalism*, immersive journalism or 360 journalism, referring to the production of news in a way that replicates either a real or imagined environment and allows users to gain first-person experiences of the events or situation described in news stories while at the same time interacting with the environment. The fundamental idea of this new form of journalism is to allow the participant, typically represented as a digital avatar, to enter a virtually recreated scenario representing the news story. Today, 360° videos can be produced

with little technical expertise, using relatively inexpensive cameras, and posted on Facebook, Instagram and/or YouTube, allowing pieces to go viral while dramatically cutting their delivery cost. Meanwhile, the cost of headsets that deliver an even more realistic experience continues to fall (Mabrook & Singer, 2019). VR journalism is gaining widespread attention since it has the potential to enhance storytelling by offering the audience experiences and environments that traditional media could only hint at; it can also encourage attitude and behavior change to a greater extent than content produced for other platforms (Karlin B, Kim, H., T., Kelly, R., Blakley J., Brenner, C., & Riley, P., 2018). It should be emphasized, though, that this new form of journalism is best used for stories in which images are lacking because access is denied or where being in the actual space deepens one's understanding of a story beyond a written narrative, photos or regular video.

Selfie Journalism

The flourishing of social media have also significantly affected the media field, leading to the rise of *selfie journalism*: the practice of sharing and, thus, distributing to a global audience highly personalized information/stories/moments/events, following the vast growing tendency for people to broadcast themselves (Maniou & Veglis, 2016). Selfie journalism could be described as a new tendency in participatory and citizens' journalism, exercised via mobile devices and based on the practice of posting self-photographs (and/or videos) in the new media and thus reproducing content, not based on isolated objects/people/images, etc., but on artefacts framing the 'author' of the photograph. These artefacts (either objects/premises and/or other people) mostly constitute elements of the author's private life, which, thus, becomes intentionally public, aiming to attract other people's attention and provoke (mainly positive) reactions (Maniou, Panagiotidis & Veglis, 2017).

Network Journalism

The proliferation of social media and the participatory and interactive characteristics of the networked digital media have paved the way for another new form of journalism that draws on the active role and collective intelligence of members of the public. In *network journalism*, professional journalists and members of the public cooperate to produce the best stories by collaborating, corroborating, correcting and ultimately distilling the essence of the story that will be told; as such, the basic characteristics of network journalism are convergence, interactivity, customization of content and hypertextuality (Deuze & Bartoel, 2001). Countless citizen journalists contribute with their reports, images, information and opinions, making it possible for the practice of journalism to broaden the scope and diversity of its sources (Van der Haak, Parks & Castells, 2012).

Drone Journalism

The flourishing of technology is also influencing the role of drones in journalism, offering opportunities for exploring new forms of storytelling and reaching new places (Figure 1). Hovering over hostile environments and surfaces of land or water unapproachable by other aircraft, drones can diminish the distance between the audience and the subject, change and enhance the viewer's perspective in a number of distinctive ways. Drones can "pan" 360 degrees like any other camera, but if they are pointing down vertically, the rotating or "orbital" shot can circle nimbly around its subject. They can "fly-through" a

restricted space such as an archway, or dramatically “reveal” with a “pedestal” shot, rising or dropping (Adams, 2018).

These small remotely-guided aircraft, formally known as unmanned aerial vehicles, have evolved into a significant newsgathering tool for journalists that can also be combined with all the other contemporary media technologies such as VR journalism and provide footage that could not be obtained in any other way. In the years to come it is expected that drones will be equipped with environmental and pollution sensors or image recognition and crowd counting software thus helping journalists scrutinize official data, verify information or reveal new stories. Drone journalism constitutes an innovative new form of journalism that brings news perspectives into the field, but still needs to overcome a complex set of ethical and legal issues related to privacy and state restrictions regarding airspace regulations.

Figure 1. A picture taken with a drone to show the effect of climate change on the second-largest lake in Bolivia (Credit: Josh Haner/The New York Times)



Robot Journalism

Artificial Intelligence, where computers learn independently over time to improve outcomes, is developing algorithms to automatically generate news from structured data without human interference. This has led to the emergence of a new form of digital journalism, widely known as *robot journalism*, automated journalism or algorithmic journalism. This form of journalism is widely integrated into newsrooms, transforming the traditional practices of gathering, producing and disseminating information. The most common fields of use are finance, sports and election results. Paired with data storage, an algorithm

retrieves appropriate data and arrays them in accordance with its intended purpose. Afterwards, it structures a narrative and creates a story using natural language. Algorithms designed by human facilitate this process. In this sense, a robot utilized in journalism can be considered a semi-autonomous system. The final step in the process of writing news stories is to publish them (Kim & Kim, 2018). Apart from generating stories based on statistics and numerical figures, robot journalism is already used for creating video or photo displays to accompany text stories, driving better content recommendations. News organizations are increasing using ‘bots’ to deliver updates to smartphones and to interact with users in social media and send them customized content. As technology develops, robot journalism is likely to move into increasingly more challenging areas. The stakes are higher, though, when those bots speak for organizations that bill themselves as trustworthy sources of information. Misleading audiences, even accidentally or for only a moment, can damage a newsroom’s credibility (Heckman, 2018).

Data Journalism

The advent of technology is also helping journalists make sense of an enormous amount of complex data, to report on them and produce comprehensive and compelling stories through engaging infographics. *Data journalism* is a new form of journalism, referring to the use of online tools and open source software to rapidly scrape, filter, analyze and visualize stories in interactive features, mobile apps and maps (Howard, 2014). This new form of journalism, combining the skills and techniques of media professionals with the data analysis and programming skills of data analysts, is mainly used in investigative stories aiming to reveal corruption and hold the powerful to account.

SOLUTIONS AND RECOMMENDATIONS

The evolution of digital technologies and social media platforms have made it possible for journalists to gather, produce and disseminate stories in an innovative, fast and interactive way. Although acceleration seems to be the dominant trend in this era of instantaneous journalism, many media professionals and academics claim that digitization has wrought havoc on investigative journalism and stress the need for slow journalism which puts the emphasis on in-depth reporting, long form storytelling and the quest for nuances and perspectives.

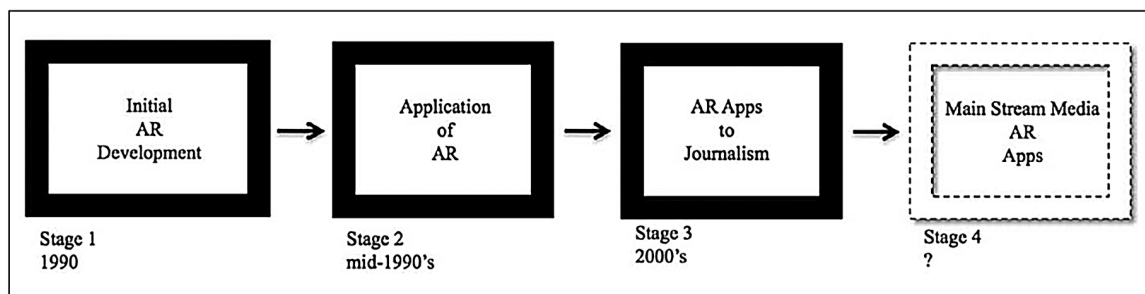
As Le Masurier (2015, p.143) points out, *slow journalism* requires time for deeper reflection and/or investigation about an original subject, avoids sensationalism, is ethical in the treatment of subjects and producers, and uses verifiable and traceable sources and transparent methodology. Instead of desperately trying to compete with social media in a 24/7 news production process with continuous deadlines, slow journalism takes the time to research, double check the information, provide context to readers and make sure that it adheres to the fundamental principles and ethical values of quality journalism. Slow journalism can embody all the innovative digital tools and social media available in order to produce the best possible result, but it goes far beyond this. By slowing down, journalists can produce long-form multimedia stories – sometimes in cooperation with colleagues from other media organizations – that allow consumers to engage with complex stories that were not told at all or gained little media attention and journalists to present more points of view and add context to complex subjects. It goes without saying that slow and investigative journalism can certainly co-exist with all the other forms of journalism, but it costs time and money, since some stories may take months of investigation. This is one of the reasons that most media organizations have turned their backs on slow journalism. On the other hand,

slow journalism offers a deeper understanding of current and important issues and offers critical thinking that many readers are willing to pay for, making this investment worthwhile. Most of all, slow journalism constitutes a creative response to the crisis of quality journalism, since it adheres to the principles of truthfulness, accuracy, objectivity, impartiality, fairness and public accountability.

FUTURE RESEARCH DIRECTIONS

Digital media are evolving in a dynamic and innovative way and it is no exaggeration to say that the current innovative forms of journalism may be considered outdated in a short period of time. Future and emerging trends in digital media and journalism are expected to evolve around smartphone technology, social media platforms and advancements in AI. As news continue to shift to a digital format, *Augmented Reality (AR) journalism*, a technology that allows a user to continue to see and hear the surrounding world but with additional sights and sounds that are synchronized to the exact location relative to a user's three-dimensional (3-D) orientation to a geographic locale, seems to be an emerging trend (Pavlik & Bridges, 2013). Using new types of glasses or contact lenses, AR journalism introduces a digital overlay into the existing reality, bringing news consumers closer to the story and enriching journalism in a new and exciting way. In their typology of the stages that AR's development is following, namely (1) initial research design and development of AR technology completely outside the realm of journalism and media field, (2) early evolution of AR into a form that could be applied to other fields of human endeavor and (3) effective utilization of AR in journalism and media organizations for several reasons. Pavlik & Bridges (2013) predict a fourth stage of development during which journalism and media organizations will fully integrate AR as a routine narrative (Figure 2), allowing the audience to 'experience' the news, rather than just reading or observing them.

Figure 2. Stages of augmented reality (AR) development (Pavlik & Bridges, 2013)



Conversational journalism, voice-powered AI that allows the consumer to interact with a bot, is another future trend. Conversational journalism is mainly based on the advent of chatbots, namely conversational bots that provide answers to questions about news events, people, and/or places. Chatbots offer a new interface for communicating with a story and help journalists tell their stories differently or collect information from readers. Their unique characteristics, such as the ability to have some kind of conversation with the user, are expected to be exploited in the near future (Veglis & Maniou, 2019). The advances in the sector of the Internet of Things (IoT), are also expected to have a serious impact on journalism as online media will extend to devices such as voice assistants, wearables, home automation objects and a whole host of elements we use in our daily lives (Salaverría, 2019).

Apart from technology, though, the need for accurate and quality journalism is also likely to pave the way for new forms of journalism that emphasize in-depth reporting and providing information and analysis that are unavailable elsewhere.

CONCLUSION

Digital media have developed rapidly in recent years and advances in technology have paved the way for the implementation of computational and automatization processes in all aspects of mass communication. Digitalization, social media platforms, smartphone technology and AI software improve the workflow speed of journalists, add value to their stories and provide consumers with innovative, engaging and in-depth storytelling. New media technologies are becoming an integral part of the digital journalist's toolbox and new forms of journalism are reshaping the media field. For example, AI robotic presenters that can imitate news anchors' voices, lip movements and expressions are already introduced by China's Xinhua News Agency. New York Times have already adopted AR in their digital storytelling extending their stories beyond the inches of a screen and offering their audience the chance to interact with augmented reality content. The next quarter century, the intersection of journalism and technology is expected to deepen and lead to major technological, social and professional innovations that will continue the profound transformation of journalism, quite possibly at an even faster pace than in recent years, generating opportunities as well as new challenges (Salaverría, 2019). To survive in this competitive environment, media organizations will need to adapt and incorporate these changes, without losing sight of their primary aim, to provide their audiences with the information and context they need to make sense of the world.

REFERENCES

- Adams, C. (2018). Tinker, Tailor, Soldier, Thief: An Investigation into the Role of Drones in Journalism. *Digital Journalism*, 1-19. Doi: doi:10.1080/21670811.2018.1533789
- Appelgren, E. (2004). Convergence and divergence in media: Different perspectives. In *ICCC 8th International Conference on Electronic Publishing 2004*, (pp. 237-248). Academic Press.
- Chadwick, A. (2013). *The Hybrid Media System: Politics and Power*. Oxford University Press. doi:10.1093/acprof:oso/9780199759477.001.0001
- Chadwick, A., Dennis, J., & Smith, A. P. (2016). Politics in the age of hybrid media: Power, systems and medialogics. In A. Bruns, G. Enli, E. Skogerbo, A. O. Larsson, & C. Christensen (Eds.), *The Routledge Companion to Social Media and Politics* (pp. 7-22). Routledge.
- Cushion, S., & Sambrook, R. (Eds.). (2016). *The Future of 24-Hour News: New Directions, New Challenges*. Peter Lang. doi:10.3726/978-1-4539-1816-6
- Deuze, M., & Bartoel, J. (2001). Network Journalism: Converging Competences of Media of Media Professionals and Professionalism. *Australian Journalism Review*, 23(2), 91-103.
- Heckman, M. (2018). Used carefully, chatbots can be an asset to newsrooms. *Columbia Journalism Review*. Available at: <https://www.cjr.org/innovations/chatbots.php>

Howard, A. B. (2014). *The art and science of data-driven journalism*. Columbia University Academic Commons., doi:10.7916/D8Q531V1

Karlin, B., Kim, H. T., Kelly, R., Blakley, J., Brenner, C., & Riley, P. (2018). Does Medium Matter? Exploring the Role of Virtual Reality in Journalism. *The Media Impact Project of the USC Annenberg Norman Lear center*. Available at <https://www.mediaimpactproject.org/frontline-vr-study.html>

Kim, D., & Kim, S. (2018). Newspaper journalists' attitudes towards robot journalism. *Telematics and Informatics*, 35(2), 340–357. doi:10.1016/j.tele.2017.12.009

Le Masurier, M. (2015). What is slow journalism? *Journalism Practice*, 9(2), 138–152. doi:10.1080/17512786.2014.916471

Mabrook, R., & Singer, J. B. (2019). Virtual Reality, 360° Video, and Journalism Studies: Conceptual Approaches to Immersive Technologies. *Journalism Studies*, 1(17), 1–17. doi:10.1080/1461670X.2019.1568203

Maniou, T., Panagiotidis, K., & Veglis, A. (2017). The politicization of selfie journalism: An empirical study to parliamentary elections. *International Journal of E-Politics*, 8(2), 1–16. doi:10.4018/IJEP.2017040101

Maniou, T., & Veglis, A. (2016). 'Selfie journalism': Current practices in digital media. *Studies in Media and Communication*, 4(1), 111–118. doi:10.11114mc.v4i1.1637

Manovich, L. (2007). Understanding hybrid media. In B. S. Hertz (Ed.), *Animated Painting* (pp. 36–45). San Diego Museum of Art.

Pavlik, J. V., & Bridges, F. (2013). The emergence of augmented reality (AR) as a storytelling medium in journalism. *Journalism & Communication Monographs*, 15(1), 4–59. doi:10.1177/1522637912470819

Report, D. F. (2018). *Annual study. Center for the Digital Future at USC Annenberg*. Available at: <https://www.digitalcenter.org/digital-future-project/>

Salaverría, R. (2019). Digital journalism: 25 years of research. Review article. *El Profesional de la Información*, 28(1), 1–26. doi:10.3145/epi.2019.ene.01

Van der Haak, B., Parks, M., & Castells, M. (2012). The future of journalism: Networked journalism. *International Journal of Communication*, 6(16), 2923–2938.

Van der Wurff, R. (2008). The effects of Internet upon media content. In L. Kung, R. Picard, & R. Towse (Eds.), *The Internet and the Mass Media* (pp. 91–115). Sage. doi:10.4135/9781446216316.n4

Veglis, A., & Maniou, T. A. (2019). Chatbots on the Rise: A New Narrative in Journalism. *Studies in Media and Communication*, 7(1), 1–6. doi:10.11114mc.v7i1.3986

Webster, J., & Ksiazek, Th. (2012). The dynamics of audience fragmentation: Public attention in the age of digital media. *Journal of Communication*, 62(1), 39–56. doi:10.1111/j.1460-2466.2011.01616.x

ADDITIONAL READING

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, 8(5), 519–531. doi:10.1080/17512786.2014.883116

Franklin, B. (2014). The future of journalism: In an age of digital media and economic uncertainty. *Journalism Studies*, 15(5), 481–199. doi:10.1080/1461670X.2014.930254

Nielsen, R. K. (Ed.). (2015). *Local Journalism: The Decline of Newspapers and the Rise of Digital Media*. I.B. Tauris. doi:10.5040/9780755695171

Pavlik, J. V., & Bridges, F. (2013). The emergence of augmented reality (AR) as a storytelling medium in journalism. *Journalism & Communication Monographs*, 15(1), 4–59. doi:10.1177/1522637912470819

Pickard, R. (2014). Twilight or new dawn of journalism? Evidence from the changing news ecosystem. *Digital Journalism*, 2(3), 273–283. doi:10.1080/21670811.2014.895531

State of the News Media. (2017). *Data and trends about key sectors in the U.S. news media industry*. Pew Research Center. Retrieved May 20, 2018 from <https://www.pewresearch.org/topics/state-of-the-news-media/2017/>

Veglis, A., & Bratsas, C. (2017). Towards a taxonomy of data journalism. *Journal of Media Critiques*, 3(11), 109–121. doi:10.17349/jmc117309

KEY TERMS AND DEFINITIONS

Conversational Journalism: A term used to refer to voice-powered Artificial Intelligence that provides answers to questions given by users.

Data Journalism: The use of software and online tools in order to analyze and visualize a large amount of complex data.

Digital Media: Digitalized information or other content that can be transmitted via the internet and/or a computer network.

Drone Journalism: The use of remotely operated or autonomous drones to acquire pictures and videos that can enrich media content.

Hybridity: A mixed system including traditional, digital, and new media.

Interactivity: A term used to distinguish new from old media, which refers to the communication process between users and between users and computers, through the media.

Network Journalism: A term used to refer to professional journalists and other stakeholders cooperating to produce stories. Networks of professionals and citizens collaborate, corroborate, correct, and ultimately distill the essence of the story that will be told.

Robot Journalism: The use of Artificial Intelligence in order to develop algorithms that can generate news and stories from structured data.

Selfie Journalism: A term used to refer to a new tendency in participatory and citizens' journalism, exercised via mobile devices and based on the practice of posting self-photographs (and/or videos) in social media and thus reproducing content, not based on isolated objects/people/images, etc., but on artefacts framing the 'author' of the photograph. These artefacts (either objects/premises or other people) mostly constitute elements of the author's private life, which, thus, become intentionally public, aiming to attract other people's attention and (mainly) positive reactions.

Slow Journalism: A term used to refer to the kind of journalism that requires time for deeper reflection and/or investigation about an original subject, avoids sensationalism, is ethical in the treatment of subjects and producers, and uses verifiable and traceable sources and transparent methodology. Instead

of desperately trying to compete with social media in a 24/7 news production process with continuous deadlines, slow journalism takes the time to research, double check the information, provide context to readers and make sure that it adheres to the fundamental principles and ethical values of quality journalism.

Smart Technology: The term ‘smart’ is in fact an abbreviation for ‘Self-Monitoring Analysis and Reporting Technology’, although it is used with its actual meaning of ‘clever’ to describe any kind of available technology that allows users to connect with internet networks.

Virtual Reality Journalism: The term refers to the production of news in a way that replicates either a real or imagined environment and allows users to interact within this environment.

Data Journalism: Definition, Skills, Difficulties, and Perspectives

Andreas Veglis

 <https://orcid.org/0000-0002-0286-2304>

Aristotle University of Thessaloniki, Greece

Charalampos Bratsas

Aristotle University of Thessaloniki, Greece

INTRODUCTION

The introduction of ICTs (Information Communication Technologies) at the last decade of the 20th century had a profound impact on almost every aspect of human activities. Journalism was one of the areas where ICTs had a significant impact, since it transformed the profession through the digitalization of the work process as well as the introduction of internet services (Veglis 2009). Today, the journalist is expected to have the ability to firstly employ many tools and services in order to be instantly informed about breaking news as well as current events, and secondly, use a variety of tools and applications in order to prepare and disseminate news articles (Veglis & Bratsas, 2017a). Many new types of journalism have emerged (algorithmic journalism, drone journalism, multimedia journalism, etc. (Diakopoulos, 2019; Ntalakas, Dimoulas, & Veglis & Bratsas, 2017a; 2017b; Tu, 2015)) along with data journalism (Gray, Chambers, & Bounegru, 2012), which require journalists to have special skills.

In recent years, data journalism has drawn significant attention in academic literature as well as in the area of new developments in digital news production (Hermida, & Young, 2017; Loosen, Reimer, & De Silva-Schmidt, 2017; Weber, Engebretsen, & Kennedy, 2018). Data journalism is now considered to be an established form of journalism. It has appeared gradually in the dawn of the new century. Many factors have contributed to the introduction of data journalism, but one of the most prominent is believed to be the availability of data in digital form. Another factor that contributed to its introduction was the availability of visualization and data management tools (Veglis & Bratsas, 2017a). Data Journalism is a journalistic specialty reflecting the increased role that numerical data has in the production and distribution of information in the digital era. Data can be the source of data journalism, and/or it can be the tool with which the story is told (Gray, Chambers, & Bounegru, 2012).

This chapter attempts to describe the state of data journalism today. The background section provides the historic evolution and definitions of data journalism. Next, the stages of data journalism are presented in detail, along with the necessary corresponding skills of the data journalist. Also, the various difficulties that the evolution of data journalism is facing are presented. Finally, recommendations and future research directions are briefly discussed.

DOI: 10.4018/978-1-7998-3479-3.ch079

BACKGROUND

Evolution and Definition

Although the term data journalism started to attract attention at the end of the previous century, initial examples of data journalism appeared quite early. According to Simon Rogers (currently data journalist and data editor at Google), the first example of data journalism was published at The Guardian in 1821, and it concerned the number of students who attended school and the costs per school in Manchester (Gray, Chambers, & Bounegru, 2012).

At the end of the 20th century, employing large datasets to write an article was difficult and required skills that went beyond the capabilities of the average journalist. To address these shortcomings news organizations in the United States and Great Britain began hiring programmers to work on novel news products (Parasie & Dagiral, 2013). Traditionally, journalists have relied on information provided by a variety of sources (governments, officials, research studies, etc.). Of course, there were some cases of investigative journalism where journalists were able to find resources to gather and analyze their own data and publish their results in articles (Veglis & Bratsas, 2017a). But things changed significantly as data gradually became more available online, along with efficient tools allowing anyone to analyze, visualize and publish large datasets (Gray, Chambers & Bounegru, 2012; Sirkkunen, 2011; Uskali & Kuutti, 2015).

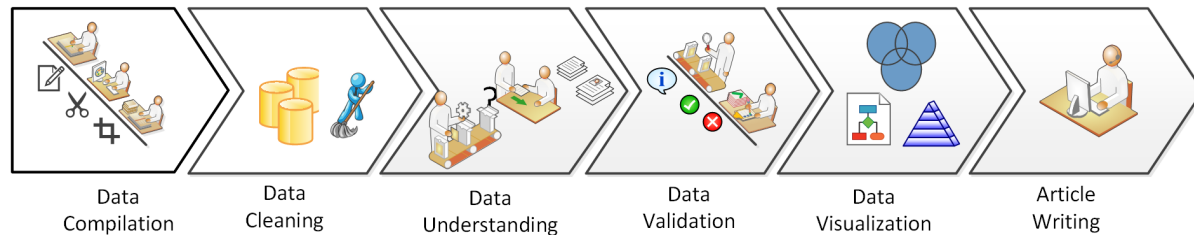
Big data, the popular term for large datasets, was introduced into common usage during the last decade of the 20th century. It can be defined as datasets of a size which cannot be captured, curated, managed and processed by commonly used software running on standard personal computers (Snijders, Matzat & Reips, 2012). Is it quite obvious that there is a strong relation between big data and data journalism, since they have developed almost in parallel and data journalism is based/generated by big data (Veglis & Maniou, 2018).

The concept of data journalism is not new. It has been around since the beginning of the digitalization of news. Digital data has been utilized in news production since the late 60s in US newspapers (Parasie & Dagiral, 2012). Data journalism gradually emerged with the rapid introduction of ICTs and the availability of data in digital form. The term *data journalism* is synonymous with *data-driven journalism*, while the older term *computer-assisted reporting* has largely vanished since it was introduced at the early stages of computer history (Bradshaw, 2010). It is worth noting that in the case of data journalism there is an increased interaction between journalists and several other fields such as design, computer science and statistics (Thibodeaux, 2011; Veglis & Bratsas, 2017a).

The term data journalism is attributed to Simon Rogers who first mentioned it in a post to the Guardian Insider Blog (Knight, 2015). It can be viewed as a process that begins with analyzing and continues with filtering and visualizing data in a form that links to a narrative (Lorenz, 2010). It combines spreadsheets, graphics, data analysis and the biggest news stories (Rogers, 2008). It is fundamentally the production of news graphics and includes elements of design and interactivity (Bradshaw, 2010; Lorenz, 2010; Rogers, 2008). Megan Knight (2015) describes data journalism as “a story whose primary source or ‘peg’ is numeric (rather than anecdotal), or a story which contains a substantial element of data or visualization”.

Veglis and Bratsas (2017a) proposed a definition in order to better address the power of visualization and interactivity which are significant factors in data journalism. They defined data journalism as the process of extracting useful information from data, writing articles based on the information and embedding visualizations (interacting in some cases) in the articles that help readers understand the significance of the story or allow them to pinpoint data that relate to them.

Figure 1. Data Journalism workflow



Data Journalism Skills

Veglis and Bratsas (2017a) organized the data journalism workflow in six stages, namely: *Data Compilation*, *Data Cleaning*, *Data Understanding*, *Data Validation*, *Data Visualization* and *Article Writing*. The workflow model is depicted in figure 1. In order for journalists to be able to produce data journalism projects they need to process certain skills that directly relate with the data journalism stages. Next, we briefly describe each stage, as well as the corresponding data journalist's skills.

Data Compilation: A data journalism project begins in one of two ways: either the journalist has a question that needs data or has a dataset that needs questioning. The compilation of data can take one of the following forms: (i) data may be supplied directly by an organization (in some cases in the form of open data), (ii) data may be found with the help of advanced searching techniques, (iii) data may be compiled by scraping web pages, (iv) data may be collected by converting documents to other formats that can be analyzed, and (v) data may be collected by means of observation, surveys, online forms or crowdsourcing (Bradshaw, 2011). Data scraping is a process in which a software tool extracts data from a human-readable output that originates from other software (i.e. copying in a spreadsheet the content of a table which is included on a webpage). Also, there are free tools on the internet that allow users to convert documents to other forms that can be analyzed. For example, DocumentCloud is a web-based software platform created specifically for journalists to allow the searching, analyzing, annotation and publication of primary source documents used in reporting (Veglis & Bratsas, 2017b).

First of all, a data journalist has to be capable of generating ideas for data-driven stories and determining the relevant data that he/she needs for his/her research. Then, he must be able to find and acquire the data he is searching for in many ways. While web scraping may be considered as a more advanced skill that often depends on the web page technologies, the awareness of open data repositories and statistical platforms and databases is a crucial part for data retrieval, as governments, public sector bodies and analytics agencies tend to provide more and more open data lately. The data journalist must have a broad knowledge of all the corresponding national open data sources, at least, and how he can use them, in order to extract the data for his analysis. He should also know to extract data from social network APIs, like Twitter, as these can provide additional data directly related to a topic, or an entity, although they usually grant access to only a small amount of the most recent data (Veglis & Bratsas, 2017a).

Data Cleaning: Also known as *data scrubbing*, this is the process of detecting and correcting corrupted or incorrect records from a dataset (Wu, 2013). This can be accomplished by removing human errors and converting the data into a format that is consistent with other data the journalist is using. Typical examples include empty entries, duplicate entries, the use of default values to save time or where no information was held, incorrect formatting (for example words instead of numbers), corrupted entries or entries with HTML code, multiple names for the same thing and missing data (Bradshaw, 2011). Cleaning

data can be done in simple ways, for example, using find and replace commands or filters in spreadsheets. There are also specialized tools, like Google's OpenRefine, which is a standalone open source desktop application for data cleanup and transformation to other formats. The application's interface resembles spreadsheet applications and it can open files of various spreadsheet formats (Veglis & Bratsas, 2017a).

Data journalism skills are rather obvious in this step. The data journalist must be able to fluently use the most common data cleaning tools, spreadsheet software and SQL. He must also have in mind all the potential errors that could be included in the extracted dataset, in order to check and correct the records in an automated way. As each dataset may contain different kinds of errors, this skill needs continuous and repetitive effort with many datasets that require various data cleaning tasks.

Data Understanding: Datasets usually include various codes that represent categories, classifications or locations, and special terminology that it is not understood by journalists. Frequently further data is required for existing data to become meaningful and the journalist has to put them in the story context. The data journalist interviews not a person but the data, as he explores his initial ideas through them, using statistical tools. Spreadsheets, relational databases and more advanced software like R, Python, Ruby, NodeXL and network analysis are some of the main tools to find the story in the data, to investigate and analyse the data, to identify significant data trends and finally, to lead to insightful results and tell a compelling story.

Overall journalists must be data-literate, meaning that they must be able to consume knowledge, produce coherently and think critically about data (Veglis & Bratsas, 2017a). Additionally, journalists must possess statistical knowledge and computational skills, think of creative ways to channel data and also understand how to work with large datasets, how they were produced, how to connect various datasets and how to interpret them (Gray, Chambers, & Bounegru, 2012).

Data Validation: This stage includes the process of cross-checking the original data and obtaining further data from sources in order to enrich the available information (Silverman, 2014; Veglis, 2013). It is worth noting that, like any source, datasets cannot always be trusted since they come with their own histories, biases, and objectives. That means that journalists have to investigate issues like: who gathered it, when, and for what purpose, and how it was gathered (Bradshaw, 2011). This can be accomplished by investigating the history of the creation of the dataset, by finding references to the dataset or by using other sources of information that refer to the same subject (Silverman, 2014; Veglis & Bratsas, 2017a).

The data journalist needs to develop ways and skills to check and prove the correctness of the data used. Thus, data journalism skills are related at some level to the data compilation step, as the journalist must be able to search for data as well as its metadata, that is information about the data (e.g. author, date produced, updated, or published, etc). Usually, open data repositories from public sector provide adequate metadata and this is primarily a data publisher's task. However, the data journalist is responsible for checking and cross-validating the datasets and using only correct and updated data in his research. At some point, the data journalist acquires the experience and the skill of identifying the reliable sources and datasets, as well as what is missed or has to be further examined by outdated or unofficial resources.

Data Visualization: Data visualization is a modern branch of descriptive statistics that involves the creation and study of the visual representation of data. It is the graphical display of abstract information for data analysis and communication purposes (Cairo, 2012). Static data visualizations offer only pre-composed "views" of data. Interactive data visualization supports multiple static views in order to present a variety of perspectives on the same information. Important stories include "hidden" data and interactive data visualization is the appropriate way to discover, understand and present these stories. In

interactive data visualization, there is a user input (a control of some aspect of the visual representation of information) and the changes made by the user must be incorporated into the visualization in a timely manner (Veglis, 2015; Veglis & Bratsas, 2017a).

It is worth mentioning that infographics are also part of the static visualization. Infographics are graphic visual representations of data or knowledge, which are able to present complex information quickly and clearly (Smiciklas 2012). They are often used in newspapers, to show the weather, as well as maps, site plans, and graphs for statistical data (Veglis & Bratsas, 2017b).

There are many types of data visualizations. Heer, Bostock, and Ogievetsky (2010) defined the types (and also their sub-categories) of data visualization: (i) Time-Series Data (Index Charts, Stacked Graphs, Small Multiples, Horizon Graphs), (ii) Statistical Distributions (Stem-and-Leaf Plots, Q-Q Plots, Scatter Plot Matrix -SPLOM, Parallel Coordinates), (iii) Maps (Flow Maps, Choropleth Maps, Graduated Symbol Maps, Cartograms), (iv) Hierarchies (Node-Link Diagrams, Adjacency Diagrams, Enclosure Diagrams) and (v) Networks (Force-Direct Layout, Arc Diagrams, Matrix Views).

Today there are a lot of available online tools that can be used for creating interactive data visualizations. All of them are either free or offer a free version (except a paid one that includes more features). Some of the most widely used tools are CartoDB, Google Chart Tools, Google Fusion Tables, Tableau Public, Many Eyes, and Infogram. A complete list of the available tools can be found at <http://selection.datavisualization.ch>.

The visualization is a crucial part of the story and its presentation. The data journalist must be able to make insightful and illustrative visualizations, that highlight information that is not clear within the raw data and incorporate them harmoniously into the story. Apart from the ability of using sufficient tools for visualizing the data to transform a dataset into a powerful and insightful visualization, the data journalist needs to know, at first, the pros and cons of different methods of visualizing data, as well as to understand the relations within the data that have to be represented by a chart or diagram. Different chart categories may be used for showing a comparison, or a proportion, or a hierarchy, or a distribution, etc. Thus, the data journalist must have a clear story objective and select only from a few charts of his visualization quiver to represent the data.

Article Writing: The last stage in a data journalism project includes the writing of the news article. Depending on the intended publication medium, the article may include special characteristics (*e.g.*, external links, other articles or related material, multimedia content, mashups, static or interactive visualizations) in order to fully take advantage of the medium's potentials (Veglis & Bratsas, 2017a). The amount of text that is included in the data journalism article along with the visualizations may vary considerably. Specifically, we can have the case where the visualization supplements the text (which is quite extended) as well as the case where the visualization is the center of the project and the text plays a supplemental role, explaining parts of the visualization.

The data journalist must be able to present all the key points of his research and integrate them properly into a news story. All ideas and information have to be well presented and understandable to the audience, who are likely coming from non-statistical background, otherwise the story will have little effect, no matter how insightful or valuable it is. Feature and multimedia storytelling are also prerequisite skills. Finally, all the stages must be clear to the audience and the methodology of the data journalist needs to be reproducible, in order to be verifiable and repeatable.

DIFICULTIES

8

Although data journalism has been around since the beginning of the 21st century, it does not appear to have gained momentum (Veglis & Bratsas, 2017a; 2017b). It is common practice for data journalism stories to be categorized as less interesting stories, resulting in being published on inside pages, or as infographics, which relatively few people look at (Stoneman, 2015), and even fewer read through. Based on the above, it is quite obvious that data journalism does not appear to be near in becoming mainstream journalism (Veglis & Maniou, 2018).

There are several factors that contribute to this result. Data journalism projects require time in order to be developed, and time is something the media organization do not have today. Also, data journalism costs more than traditional journalism. The economic model of practicing data journalism is not yet quite defined (Gray, Chambers & Bounegru, 2012), and media organizations seem reluctant to invest on data journalism without seeing tangible circulation gains.

There is also the issue of the additional skills that journalists must process in order to be able to develop data journalism projects. The journalists seem quite unprepared in many cases for such tasks and expect media organizations to cover the cost of acquiring data journalism skills (Veglis & Bratsas, 2017a). Of course, young journalists are not expected to encounter similar difficulties (at least not to the same extent) since market demands, as far as journalistic skills are concerned, has forced journalism educators to adjust their programs in order to better facilitate the needs of the industry (Wenger & Owens, 2012). Currently, data journalism is taught in a limited number of journalism schools in Europe. Of course, this is something worth investigating more thoroughly in the near future. A recent survey, concerning the use of data journalism in Greece found a low penetration of data journalism practices in media organizations. Nevertheless, the majority of the journalists appeared to agree with the importance of working with data and was interested in acquiring more knowledge and skills concerning data journalism (Veglis & Bratsas, 2017a).

FUTURE RESEARCH DIRECTIONS

Tim Berners-Lee believes that “data-driven journalism is the future” and urges reporters to hunt for stories in datasets (Arthur, 2010). We are convinced that the Semantic Web and Linked Open Data will play a significant role in the evolution of data journalism. In the future, journalists should be able to comprehend and utilize advanced technologies which will include a more “intelligent web” (Bradshaw & Rohumaa, 2011). Semantic Web technologies, smart devices (smartphones etc.) and tools are continuously being transformed and upgraded. Data journalism ought to adopt these technologies. It is worth noting that Web 3.0 and open data can play an important role in the case of data journalism as they support the acquisition and validation of data which are the main source of data journalism.

One significant issue that needs to be addressed is the lack of online platforms that can support all stages of a data journalism project. Today, there are several online tools that can be used in different stages of a data journalism project. Each tool has its own unique interface and the journalist must spend time in order to learn how to use it effectively. Most of them do not offer any interconnectivity with other tools. Thus, the journalist spends a lot of time transferring data from one platform to another in order to exploit the strong points of each tool. In other words, the landscape is fragmented, and there is a lack of an integrated platform on which the journalists can work on all stages of data journalism projects. Data journalists can strongly benefit from the introduction of online platforms that will interface various

online tools that can be used in data journalism projects. This is a problem that media industry should address in cooperation with the informatics industry.

CONCLUSION

This chapter discusses the issue of data journalism. Data journalism is considered to be a journalistic specialty that is expected to flourish in the world of data that we are living in today. Data journalism stages were presented and discussed in detail, in conjunction with the skills that journalist ought to have in order to develop data journalism projects. Also, the difficulties that data journalism is facing are also examined. Finally, future perspective of this journalism specialty are also briefly discussed.

There is no doubt that the majority of the journalists today are not involved in data journalism projects, although they are willing to work on such projects in the future. Online journalists appear to be more involved and more motivated towards data journalism. Since there is an ongoing shift to online journalism (Cokley, Edstrom, McBride, & Ranke, 2015), many new developments in the area of data journalism are expected. In the continuously changing media landscape, where traditional media organizations are faced with stiff competition from social networking services as well as from citizen journalists, the degree of complexity that data journalism involves in its production may be an opportunity for professional journalists to gain back their popularity as well as their credibility.

REFERENCES

- Arthur, C. (2010, November 22). Analysing data is the future for journalists, says Tim Berners-Lee. *The Guardian*. Retrieved March 18, 2016, from <https://www.theguardian.com/media/2010/nov/22/data-analysis-tim-berners-lee>
- Bradshaw, P. (2010, October 1). How to be a data journalist. *The Guardian Data Journalism*. Retrieved March 18, 2016, from <http://www.Guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide>
- Bradshaw, P. (2011). The inverted pyramid of data journalism. *Online Journalism Blog*. Retrieved March 18, 2016, from <http://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism>
- Bradshaw, P., & Rohumaa, L. (2011). *The Online Journalism Handbook: Skills to survive and thrive in the digital age*. Longman Publishing Group.
- Cairo, A. (2012). *The Functional Art: An introduction to information graphics and visualization* (Voices That Matter). New Riders.
- Cokley, J., Edstrom, M., McBride, J., & Ranke, A. (2011). Moving Away from 'Big Media': Students, Jobs and Long-tail Theory. *Australian Journalism Review*, 33(1), 107–118.
- Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press. doi:10.4159/9780674239302
- Gray, J., Chambers, L., & Bounegru, L. (2012). *The Data Journalism Handbook*. Retrieved March 18, 2016, from <http://datajournalismhandbook.org>

- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Magazine Communications of the ACM*, 53(6), 59–67. doi:10.1145/1743546.1743567
- Hermida, A., & Young, M. L. (2017). Finding the data unicorn: A hierarchy of hybridity in data and computational journalism. *Digital Journalism*, 5(2), 159–176. doi:10.1080/21670811.2016.1162663
- Knight, M. (2015). Data Journalism in the UK: A preliminary analysis of form and content. *Journal of Media Practice*, 16(1), 55–72. doi:10.1080/14682753.2015.1015801
- Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2017). Data-driven reporting: An on-going (r) evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. *Journalism*. doi:10.1177/1464884917735691
- Lorenz, M. (2010). *Data driven journalism: What is there to learn?* Paper presented at IJ-7 Innovation Journalism Conference, Stanford, CA.
- Ntalakas, A., Dimoulas, C. A., Kalliris, G., & Veglis, A. (2017). Drone journalism: Generating immersive experiences. *Journal of Media Critiques*, 3(11), 187–199. doi:10.17349/jmc117317
- Parasie, S., & Dagiral, E. (2013). ‘Data-driven journalism and the public good: “Computer-assisted-reporters” and “programmer-journalists” in Chicago’. *New Media & Society*, 15(6), 853–871. doi:10.1177/1461444812463345
- Rogers, S. (2008, December 18). Turning official figures into understandable graphics, at the press of a button. *Inside the Guardian Blog*. Retrieved March 18, 2016, from <https://www.theguardian.com/help/insideguardian/2008/dec/18/unemploymentdata>
- Silverman, C. (2014). *Verification Handbook*. European Journalism Centre. Retrieved March 18, 2016, from <http://verificationhandbook.com/downloads/verification.handbook.pdf>
- Sirkkunen, E. (2011). *Trends in Data Journalism*. Wp 3, Hyperlocal D.3.2.1.2.b. Retrieved March 18, 2016, from http://virtual.vtt.fi/virtual/nextmedia/Deliverables-2011/D3.2.1.2.B_Hyperlocal_Trends_In%20Data_Journalism.pdf
- Smiciklas, M. (2012). *The Power of Infographics: Using Pictures to Communicate and Connect with Your Audience*. Pearson Education Inc.
- Snijders, C., Matzat, U., & Reips, U. D. (2012). Big data: Big gaps of knowledge in the field of internet. *International Journal of Internet Science*, 7, 1–5.
- Stoneman, J. (2015). Does open data need journalism? Working Paper. *Reuters Institute for the Study of Journalism*. <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/Stoneman%2520-%2520Does%2520Open%2520Data%2520need%2520Journalism.pdf>
- Thibodeaux, T. (2011, October 6). 5 tips for getting started in data journalism. *Poynter: A Global Leader in Journalism*. Retrieved March 18, 2016, from <https://www.poynter.org/news/media-innovation/147734/5-tips-for-getting-started-in-data-journalism/>
- Tu, D. L. (2015). *Feature and narrative storytelling for multimedia journalists*. Focal Press. doi:10.4324/9781315851297

- Uskali, T., & Kuutti, H. (2015). Models and streams of data journalism. *The Journal of Media Innovations*, 2(1), 77–88. doi:10.5617/jmi.v2i1.882
- Veglis, A. (2013). Education of Journalists on ICTs: Issues and Opportunities. *Journal of Applied Journalism & Media Studies*, 2(2), 265–279. doi:10.1386/ajms.2.2.265_1
- Veglis, A. (2015). *Teaching Data Journalism in the School of Journalism & MC-Greece*. Paper presented at Ejta2015, Narrative Journalism and Storytelling on Digital Platforms (Teacher Conference), Milan, Italy.
- Veglis, A., & Bratsas, C. (2017a). Journalists in the age of Data Journalism: The case of Greece. *Journal of Applied Journalism & Media Studies*, 6(2), 225–244. doi:10.1386/ajms.6.2.225_1
- Veglis, A., & Bratsas, C. (2017b). Towards a taxonomy of data journalism. *Journal of Media Critiques*, 3(11), 109–121. doi:10.17349/jmc117309
- Veglis, A., & Maniou, T. A. (2018). The Mediated Data Model of Communication Flow: Big Data and Data Journalism. *KOME: An International Journal of Pure Communication Inquiry*, 6(2), 32–43. doi:10.17646/KOME.2018.23
- Weber, W., Engebretsen, M., & Kennedy, H. (2018). Data stories: Rethinking journalistic storytelling in the context of data journalism. *SComS Studies in Communication Sciences*, 2018(1), 191–206.
- Wenger, D. H., & Owens, L. C. (2012). Help Wanted 2010: An Examination of New Media Skills Required by Top U.S. News Companies. *Journalism & Mass Communication Educator*, 67(1), 9–25. doi:10.1177/1077695811426826
- Wu, S. (2013). A review on coarse warranty data and analysis. *Reliability Engineering and System*, 114, 1–11. doi:10.1016/j.res.2012.12.021

ADDITIONAL READING

- Allemang, D., & Hendler, J. (2011). *Semantic Web for the Working Ontologist* (2nd ed.). Morgan Kaufmann.
- Appelgren, E., & Nygren, G. (2014). Data Journalism in Sweden Introducing new methods and genres of journalism into “old” organizations. *Digital Journalism*, 2(3), 394–405. doi:10.1080/21670811.2014.884344
- Auer, S. R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web. Lecture Notes in Computer Science*, 4825, 722–735. doi:10.1007/978-3-540-76298-0_52
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22. doi:10.4018/jswis.2009081901
- Bradshaw, P. (2011, September 22). Data journalism survey: a mixed picture. *Data Driven Journalism*. Retrieved March 18, 2016, from http://datadrivenjournalism.net/news_and_analysis/data_journalism_survey_analysis
- Brickley, D., & Guha, R. V. (2014). RDF Shcema. Retrieved March 18, 2016, from <http://www.w3.org/TR/2014/REC-rdf-schema-20140225>

- Bull, A. (2010). *Multimedia Journalism: A Practical Guide*. Routledge. doi:10.4324/9780203866030
- Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann.
- Cheruiyot, D., Baack, S., & Ferrer-Conill, R. (2019). Data Journalism Beyond Legacy Media: The case of African and European Civic Technology Organizations. *Digital Journalism*, 1-15.
- Cokley, J., Gilbert, L., Jovic, L., & Hanrick, P. (2016). Growth of 'Long Tail' in Australian journalism supports new engaging approach to audiences. *Continuum (Perth)*, 30(1), 58–74. Advance online publication. doi:10.1080/10304312.2015.1099152
- Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., & Zijlstra, T. (2012). *The Open Data Handbook*. An Open Knowledge Foundation project, Retrieved March 18, 2016, from <http://opendatahandbook.org>
- Engelbrechtsen, M., Kennedy, H., & Weber, W. (2018). Data Visualization in Scandinavian Newsrooms: Emerging Trends in Journalistic Visualization Practices. *Nordicom Review*, 39(2), 3–18. doi:10.2478/nor-2018-0007
- European Journalism Centre. (2010). Data-driven journalism: what is there to learn? Retrieved March 18, 2016, from http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf
- Fink, K., & Anderson, C. W. (2015). Data Journalism in the United States. *Journalism Studies*, 16(4), 467–481. Advance online publication. doi:10.1080/1461670X.2014.939852
- Friedman, V. (2007, August 2). Data Visualization: Modern Approaches. *Smashing Magazine*. Retrieved March 18, 2016, from <http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches>
- Friedman, V. (2008, January 14). Data Visualization and Infographics. *Smashing Magazine*. Retrieved March 18, 2016, from <http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics>
- Friendly, M. (2007). A Brief History of Data Visualization. in C. Chen, W. Hardle and A. Unwin (eds.) *Handbook of Computational Statistics: Data Visualization*, III, Springer. Retrieved March 18, 2016, from <http://www.datavis.ca/papers/hbook.pdf>
- Mair, J., & Keeble, R. L. (2014). *Data Journalism: Mapping the future*. Abramis Academic Publishing.
- McGuinness, D. L., & van Harmelen, F. (2014). Ontology Web Language OWL. Retrieved March 18, 2016, from <http://www.w3.org/TR/owl-features>
- Murray, S. (2013). *Interactive Data Visualization for the Web*. O'Reilly.
- Open Knowledge. (n.d.). Open Data Handbook. Retrieved March 18, 2016, from <http://opendatahandbook.org/>
- Prud'hommeaux, E., & Seaborne, A. (2013). SPARQL Query Language for RDF. Retrieved March 18, 2016, from <http://www.w3.org/TR/rdf-sparql-query>
- Sandoval-Martín, M. T., & La-Rosa, L. (2018). *Big Data as a differentiating sociocultural element of data journalism: the perception of data journalists and experts*.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In Proceedings at the *IEEE Symposium on Visual Languages*, 336 – 343. 10.1109/VL.1996.545307

Spyridou, L. P., Matsiola, M., Veglis, A., Kalliris, G., & Dimoulas, C. (2013). Journalism in a State of Flux: Changing Journalistic Practices in the Greek Newsroom. *The International Communication Gazette*, 75(1), 76–98. doi:10.1177/1748048512461763

Stabellini, B., Remondino, C. L., & Tamborrini, P. (2017). Data Visualization Collection. How graphical representation can inspect and communicate sustainability through Systemic Design. *The Design Journal*, 20(sup1), S1673-S1681.

Stalph, F., & Borges-Rey, E. (2018). Data journalism sustainability: An outlook on the future of data-driven reporting. *Digital Journalism*, 6(8), 1078–1089. doi:10.1080/21670811.2018.1503060

Vatant, B., & Vandenbussche, P. Y. (2014). *Linked Open Data Vocabularies*, Retrieved March 18, 2016, from <http://lov.okfn.org/dataset/lov>

Veglis, A., & Pomportsis, A. (2012). The e-citizen in the cyberspace – A journalism aspect. In *Proceedings of the 25th International Conference on Information Law and Ethics*. Kerkyra, Greece: Nomiki Bibliothiki.

Veglis, A., & Pomportsis, A. (2014). Journalists in the age of ICTs: Work demands and Educational needs. *Journalism & Mass Communication Educator*, 69(1), 61–75. doi:10.1177/1077695813513766

Ward, M., Grinstein, G., & Keim, D. (2010). *Interactive Data Visualization: Foundations, Techniques, and Applications*. A K Peters/CRC Press.

Ware, C. (2008). *Visual Thinking: for Design*. Morgan Kaufmann Series in Interactive Technologies.

Ware, C. (2012). Information Visualization. Third Edition: Perception for Design.” Morgan Kaufmann.

Young, M. L., Hermida, A., & Fulda, J. (2018). What makes for great data journalism? A content analysis of data journalism awards finalists 2012–2015. *Journalism Practice*, 12(1), 115–135. doi:10.1080/17512786.2016.1270171

Zamith, R. (2019). Transparency, interactivity, diversity, and information provenance in everyday data journalism. *Digital Journalism*, 1-20.

KEY TERMS AND DEFINITIONS

Data Cleaning or Data Scrubbing: The process of detecting and correcting corrupted or incorrect records from a dataset.

Data Journalism: The process of extracting useful information from data, writing articles based on the information and embedding visualizations in the articles that help readers understand the significant of the story or allow them to pinpoint data that relate to them.

Data Scraping: The process in which a software tool extracts data from human-readable output that originates from other software.

Data Validation: The process of cross-checking the original data and obtaining further data from sources in order to enrich the available information.

Data Visualization: The graphical display of abstract information for data analysis and communication purposes.

Dataset: Is a collection of data that contains individual data units organized in a specific way and accessed by a specific access method that is based on the dataset organization.

Open Data: Is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike.

Exploring the Integration of User-Generated Content in Media Organizations Through Participatory Journalism

Theodora A. Saridou

Aristotle University of Thessaloniki, Greece

Andreas Veglis

 <https://orcid.org/0000-0002-0286-2304>

Aristotle University of Thessaloniki, Greece

INTRODUCTION

The concept of user engagement in news production process has evolved during the last centuries to extend from early forms of participation in printed matters to wholly participatory platforms hosted by mainstream online media outlets. The initial steps of the 17th century (Boczkowski, 2005) were followed by practices like letters to the editor, newspaper ombudsmen and reader representatives, before Internet and World Wide Web (Web) spread made it possible for message boards and chat rooms to be implemented on news websites (Deuze, 2006; Gillmor, 2004; Singer et al., 2011). Nevertheless, it was the predominance of Web 2.0 that led to the radical transformation of the journalist-audience relationship (Reader, 2018; Sinton, 2018).

Remarkable quantities of user-generated content (UGC), basically in the form of text, photos, videos and graphics (Anderson, 2007), are produced and uploaded thanks to easy-to-use publishing tools, always-on connections and powerful mobile devices (Bowman & Willis, 2003). Meanwhile, citizen journalism activity through social media such as Facebook and Twitter is becoming increasingly relevant and important (Kim & Lowrey, 2015, p. 298) and anyone with something engaging to share can be a potential mass communicator (Moore & Hatcher, 2018). As Newman (2009) states, social media and UGC have fundamentally changed the nature of breaking news by contributing to the compression of the “news cycle”, putting more pressure on editors over what and when to report and creating an important extra layer of information and diverse opinion.

Technological developments, along with social and economic transitions (Moretzsohn, 2006), declined audience’s confidence in mainstream journalism (Hirst, 2011) and loss of advertising revenues (Franklin, 2014) provoke strategic redesign in media, challenging traditional concepts such as gatekeeping (Bruns, 2005). Thus, online media adopt tools and applications of participatory journalism, where users can -among others- comment on published stories, send their own digital material, discuss several issues and share articles in social networking sites. However, users’ involvement in making and disseminating news raises a variety of ethical, legal and responsibility issues, which force organizations to use quality assurance methods (Saridou & Veglis, 2016, p. 88).

This chapter aims to offer an in-depth description of the concept of participatory journalism, by providing a deep approach to the conceptual sense of this phenomenon, presenting an analytical framework of the audience engagement in online news production. The chapter seeks better understanding

DOI: 10.4018/978-1-7998-3479-3.ch080

of the adoption of UGC in media via different forms, tools and applications and during different stages of journalistic process. The problems organizations have to deal with when amateur content is involved with the professional content and the solutions chosen are investigated as well. Finally, future research perspectives are discussed.

BACKGROUND

Very early stages of audience's participation in news production process are rooted in the 17th century, when readers can write on the intentionally blank last page of "*Publick Occurrences Both Foreign and Domestick*", before passing it to someone else (Boczkowski, 2005, p. 141). During the next century, newspapers publish readers' letters to the editor, while later, in the 20th century, more active contact with the community takes place, via newspaper ombudsmen, reader representatives, focus groups and polls (Bowman & Willis, 2003; Deuze, 2006). Apart from press, both in radio and in television broadcasting, presenters receive phone-ins and text messages from audience members who provide information about breaking news or discuss on air (Gillmor, 2004). According to Nip (2006, p. 214), it is the public journalism movement in 1990's that aims to help democracy by connecting to the community, engaging individuals as citizens and helping public deliberation in search for solutions. At the same time, newspapers, which have already launched their online media counterparts, invite users to take part in discussion board and chat room conversations (Beyers, 2004). What's more, journalists' emails are often provided next to the stories and readers use them to indicate errors or make suggestions for further research.

With the proliferation of Information and Communication Technologies (ICT), technology determines the functional definition of a new form of journalism (Deuze, 2003). As Siapera and Veglis (2012) note, online journalism causes such fundamental and permanent changes on the face of journalism that it is difficult to imagine journalism being practiced exclusively outside Internet. Journalistic practice moves towards interactivity (Stovall, 2004), individuals look for and produce additional information (Singer et al., 2011) and the way they communicate with each other is revolutionized (Hess, 2013). Ward (2002) distinguishes three levels of interactivity at which consumers interact both with the news provider and with each other, while they also have the ability to become providers.

Except for constantly evolving technological developments, broader citizens' engagement in news production process is also favored by the conditions of the economic and social environment (Domingo et al., 2008; Paulussen, Heinonen, Domingo, & Quandt, 2007). As Hirst (2011, p. 15) notices, society believes that journalists fail in their duty to promote and protect public interest and, in such a context, part of the audience seeks for alternative forms of information outside traditional media and without professional mediation (Turner, 2010). In the years that followed, breaking events, such as the 9/11 attacks in the United States of America and the 7/7 attacks in London, triggered citizen journalism activities, motivating eye-witness amateurs to collect and upload photos and personal narratives from the places of interest on blogs and forums. Instead of being primarily journalistic-centered, online news is increasingly user-centered, so that citizens have a much greater input on it, with a qualitative leap in the presentation of their own stories, opinions and audiovisual material (García-Avilés, 2010, p. 251).

Gradually, media landscape faces a catalytic transformation under the prevalence of Web 2.0 as the average citizen can now create media content (Meraz, 2009, p. 123), in many forms, using quite plain portable devices, tools and applications. Audience members acquire the opportunity to take actively part in the production and dissemination of news. Literature shapes new concepts in order to describe professionals' and non-professionals' changing role in news process. Deuze (2003) argues that users

are prosumers (merging the terms *producer* and *consumer*), while Bruns (2005) focuses on produsage, where participants are simultaneously *producers* and *users*, working on the collaborative and continuous building of existing content. As a result, journalists' hegemony as gatekeepers seems to be threatened, since news can now be recorded not only by controlling information flow, but by harnessing the collective knowledge of a community as well (Bruns, 2008, p. 176). Gatekeeping turns to gatewatching, considering that journalists do not define, but watch the information flow stemming from a variety of sources, receive UGC and construct a final product (Bruns, 2008). In a similar vein, Boczkowski (2005) refers to gateopening, when users actively co-develop website content and traditional work routines are questioned.

Although the idea of mediated participation is part of the long-standing debate between cyber-optimists and cyber-pessimists (Spyridou, 2018, p. 2), it is accepted that the current new media era has seen a proliferation of UGC and an increasing participation by Internet users in public communication and newsrooms seemingly are keen to integrate UGC into ordinary news practice (Tong, 2015). Paulussen and Ugille (2008) argue that professionals are forced to share control of the news production process with users, as they are increasingly involved in creating content. Thus organizations are gradually redefining their work strategies, combining amateur content production with editorial control under a professional umbrella (Deuze, 2006).

"Participatory journalism" is the concept that best describes "the processes and effects of ordinary citizens' contributions to gathering, selecting, publishing, distributing, commenting on and publicly discussing the news that is contained within an institutional media product" (Singer et al., 2011, p. 15). In Siaper's view (2012), it is constituted a distinct species of journalism due to three elements: The first is readers' direct participation in the journalistic process, as they are provided the means to post their news articles, photos and opinions, so they are or can potentially become journalists themselves. Secondly, the news is organized on the basis of importance to those running the site rather than confronting to standardized values and the last one identifies an involved and overtly partisan stance toward the political process. The latter means that, because of the reliance of this journalistic form on readers than journalists, its stance vis-à-vis the political process is not the objective or at least detached one found in other, more mainstream forms of professional journalism.

Although the interchangeable use of the terms "participatory" and "citizen" is not rare (Bowman & Willis, 2003; Gillmor, 2006; Lewis, Kaufhold, & Lasorsa, 2010), it is however important to distinguish between the two forms of journalism, mainly based on the concept of control. Participatory journalism encompasses users' contribution in the news-making process, which is solicited within a frame designed by professionals, while in citizen journalism people are responsible for gathering content, visioning, producing and publishing the news product, but professionals are not involved at all (Nip, 2006). Therefore, participatory journalism takes place when content is co-created by journalists and users on an editorially controlled platform and when the content created by users elsewhere is then acquired and used by professional media outlets.

EXPLORING PARTICIPATORY JOURNALISM

Tools and Applications of Participation

News organizations adopt a range of tools and applications to integrate users' contribution into their websites. Participatory formats have not been classified in a single typology, while research has pointed

out that there is diversity not only in the way journalists manage participation, but also throughout news production process and in the effort needed by participants in order to engage.

Singer et al. (2011) focused on the opportunities that are provided to audience members in order to get involved in the different stages of news production process. Their study of leading newspapers and their respective websites in ten Western democracies indicates that in the initial information-gathering stage, named access/observation stage, organizations are reluctant to allow readers to set the news agenda, limiting the user's role to serving as a source of information or raw audiovisual content. Journalists mainly enable users to submit photos and video in order to receive material that they are unable to get in other way (Mast & Hanegreefs, 2015), while an almost complete lack of witness contributions is also recorded in some media organizations (Saridou & Veglis, 2016).

Moving to the selection/filtering stage, Singer et al. (2011) concluded that it is the most closed of all, as newspapers do not allow readers to take part in the gatekeeping process and decide what should be the main news product. However, in rare cases in which users are allowed a semblance of control over the selection and filtering of news, it takes place in areas separated from the main website, though supervised by the parent organization. Rather closed is the processing/editing stage as well, with only a few options for citizen stories, always subject to strict editorial control. When users can participate in content production, their engagement is limited to lifestyle topics, treated as news tips than stories and explicitly distinguished from professional job (Johnson & Dade, 2018). In the distribution stage, participation seems to be the limited involvement in the dissemination of news stories, performed usually through automated content ranking features and social media sharing options. According to Singer et al. (2011), it is the interpretation stage that offers most opportunities for participation, from simple tools, such as polls, to more engaging and effort requiring ones, such as forums and comments. Users are allowed to discuss the news content produced by professional journalists, who control and moderate the submissions.

Following a different categorization, Franquet, Villa and Bergillos (2011) identified four groups of tools that apply to different propositions within the communication process. These are discussion tools, which encourage debate and audiences can discuss, compare opinions and exchange ideas about the information; socializing tools for sharing content, where users' contributions are not put onto the news websites, as services operate outside the organization and usage conditions do not seem to be defined by it; contribution tools for acquiring public's images, texts and audio; and rating tools for evaluating information.

Suau and Masip (2014) designed a typology that would allow the creation of different models of media participation, choosing as starting point the concept of interactivity, which is divided into different types used to group the interactive features. Firstly, they define a group of features and services labelled "selective interactivity", which allows users to interact with the system in order to adapt the content according to their preferences and includes mechanisms for personalization. A second group of tools, under the term "participative interactivity", takes place in the context of a user-user or user-professional relationship and enables users to interact with journalists and other users. Parameters are previously established by the medium's website. Lastly, a third group of interactive features, labelled "productive interaction", also takes place in the context of a user-user or user-professional relationship and aims to contribution of original content.

Identifying the most prevalent and widespread participatory tools, Spyridou (2018) concluded to the following nine: i) Content rating, which is rating of content using the media organization's designated system, ii) polls, which are topical questions posed by journalists and answered by users using pre-determined responses, iii) sharing through social networks via buttons provided, iv) audience footage in

the form of audiovisual material, v) collaborative content (e.g. contributing interview questions), vi) comments regarding a news item, typically submitted through a form at the bottom of a webpage, vii) discussion forums, which are discussions led by journalists with topical questions posed by the newsroom in which submissions are fully or reactively moderated or places where readers can engage in threaded online conversations or debates, viii) submission of textual material and ix) citizen blogs, created by users and hosted within the news website.

Moreover, Spyridou (2018, p. 6) noticed that, although all dimensions are continuous and significant variability exists within each category, the aforementioned participatory tools can be split into categories of participatory activity. They range from very low, which enable little control, require minimal effort and are invited post hoc to high, which require most effort and allow for most control over content as planned and created before publication. Between them there is the category of low participatory activity, which also requires some effort, but enables more control than the first one over the published content, in the sense that it will either be published wholesale or not, again as a post hoc contribution and the category of medium participatory activity, which requires some effort and enables limited control over the content that will be published. Sharing is treated as a unique category, since, although it does not require substantial creative effort, it places professional work at a new, different to the initial, public space.

Problems and Quality Assurance Methods

The incorporation of UGC in the journalistic process is not an easy task, as the coexistence of professional and amateur content raises significant ethical, legal and responsibility issues (Saridou & Veglis, 2016, p. 88; Veglis, 2013). Concerns regarding UGC have been raised about the actual practice and its implications for culture, society and politics (Manosevitch, 2011). Some journalists see comments as a necessary evil which may attract audiences and help to the financial survival of the profession, but holds intolerable standards of expression (Singer et al., 2011). Moreover, there is the perception that comments frequently contain defamation, incitement, abusive content, racism or hate speech and thus the greatest challenge posed stems from their low quality (Singer et al., 2011, p. 103). Many also point to excessive use of inappropriate language, flouting, stereotyping and superficial discourse that impede constructive public discourse (Manosevitch, 2011, p. 425).

In the field of intellectual property, Deliyanni (2012, p. 324) classifies possible problems in three broader categories: Issues concerning the protection of blogs, profiles in social networking sites and websites along with the original content posted on them, issues caused by the violation of intellectual rights when creators' work is offered for free although made for commercial exploitation and issues of consolidation and reconstitution of intellectual works intended for commercial exploitation. Generally, journalists try to maintain high quality of spelling and grammar, to choose material suitable for publication and to ensure the balance and decency of the website (Thurman, 2008). Furthermore, responsibility issues become complicated as traditional legal liability models are questioned in the light of changing author, editor and publisher's role (Valcke & Lenaerts, 2010). When users produce content on a professional website, the debate on liability concerns not only them, but the journalist, the owner of the website and even the Internet service provider as well.

In order to avoid problems, media organizations have to manage participatory spaces, using methods of control on users' activity. Decisions on the practices employed by journalists are affected by several factors such as story topics, journalistic culture, newsrooms norms, perceived interest value and website traffic goals (Singer et al., 2011). When news organizations use pre-moderation, journalists try to assess every piece of UGC before published, following the logic of traditional participatory spaces; this prac-

tice offers a high level of security, but demands high financial, human and time resources (Singer et al., 2011). On the other hand, post moderation is a more open approach, comments are first published and professionals intervene if a reason occurs; such techniques are nearly always accompanied by mandatory prior registration and users have to submit personal information in order to be accredited as commenters (Singer et al., 2011, p. 109).

When distributed moderation is chosen, users are involved in the control process, often earning some kind of privileges (Lampe & Resnick, 2004). At the same time, some media organizations use technical methods that require less human intervention. Automated moderation, for example, is a purely technical method and UGC is accepted or rejected under predefined rules and filters that detect and replace banned words, phrases, IP addresses or whole paragraphs (Veglis, 2014), whilst Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) is employed in the websites more rarely (Saridou & Veglis, 2016), aiming to block computer-derived mass entries. However, it appears that each strategy, whether autonomously or in combination with others, has an impact on how UGC is treated, not without financial and editorial stakes concerning cost, staff, time, quality and legality.

SOLUTIONS AND RECOMMENDATIONS

Apart from UGC that is produced on websites of media organizations, users also create content on social media platforms, extending news gathering process outside traditional spaces. Journalists use social media to look for breaking news events, find ideas for stories, keep in touch with their audience and gather information (Weaver & Willnat, 2016, p. 853). At the same time, however, digital content can be processed, intentionally altered or falsified and redistributed relatively easy, posing threats to mass communication and journalistic processes that can be associated to unwanted content tampering, construction of fake evidences, sharing and propagation of untrue stories (Katsaounidou & Dimoulas, 2018).

In order to exploit and verify UGC, media organizations utilize tools and platforms which enable them to obtain, sort and disseminate news (Katsaounidou, Dimoulas, & Veglis, 2018). Except for traditional methods, part of the collection, management and validation methods is sometimes based on semantic web services. However, such methods are not widely adopted by professionals and the need to trace in short time huge amounts of information coming from various sources remains crucial to the journalistic workflow (Saridou, Panagiotidis, Tsipas, & Veglis, 2018).

It is therefore important to design services which would aim at UGC exploitation, both of this submitted to media websites and of this created at social media platforms. It should be based on fully- or semi-automated methods, consisted of a content discovery procedure that would be followed by correction and moderation processes. At this point, a basic categorization would already have been developed and contributed to the initial handling of large-scale unstructured data. Afterwards, the most helpful step would be the implementation of semantic annotation tools aiming at effective management of UGC in order to prepare it for publication.

It is the role of journalist, however, that should define the structure and operation of the whole service. Although based on automation techniques, such an initiative should engage professionals not only at the stage of design, but also at the evaluation one. As news becomes more social, journalists have to develop their skills and use digital tools in the service of tracing information, forensic examination, UGC dissemination and verification (Johnston, 2016; Veglis, 2013). In order to verify content, professionals often use web-based platforms -like Truly Media (www.truly.media)- or tools developed in projects -like

Reveal (www.revealproject.eu)- which aim at finding and organizing information produced in social media and elsewhere. It thus seems inevitable that journalists' participation would enrich the procedure of designing a service for UGC management.

FUTURE RESEARCH DIRECTIONS

Future studies could focus on the stages of the creative procedure, starting from the model introduction and ending to the practical implementation of the designed UGC management service. Models that will be designed should effectively combine technical processes with professional norms, as at times of economic uncertainty, hyper-competition and diminishing accountability levels, convergence is used as a cost-effective strategy fostering low-cost and spreadable news production (Saridou, Spyridou, & Veglis, 2017).

Moreover, it is obvious that semantic Web tools can have an important impact upon how UGC may be treated by journalists. Semantic Web technologies, smart devices and tools are continuously being transformed and upgraded (Veglis & Bratsas, 2018) and so, in the future, journalists should be able to comprehend and utilize advanced technologies which will include more “intelligent web” (Bradshaw & Rohumaa, 2011). Research interest can thus focus on the way professionals are educated on the use of semantic technologies, while the study of their engagement in designing and developing UGC-focused tools can offer an insight on the integration of journalistic culture into technical procedures. In addition to this, their participation in the evaluation stage can improve the final product, especially in terms of usability and time saving. Notwithstanding the above, future studies should not neglect users' perceptions of involvement in news production in order to achieve accuracy.

CONCLUSION

This chapter examines the concept of participatory journalism, which evolves under technological and socio-economic developments, forcing media organizations to adjust their operation to the new challenges. Journalism websites offer a wide range of tools for UGC acquisition whereas at the same time huge amounts of information are created and shared by users on social media platforms. In such a context, a world of unstructured data calls for integration into the hectic pace of news production process. Hence, professionals try to incorporate UGC in their daily work routines not without facing possible threats to legitimacy and the established modus operandi, finding themselves in need of quality assurance methods.

At the same time, although semantic technologies can pave the way for the simplification of UGC exploitation, such practices are not largely spread apart from several heterogeneous tools and applications. Thus, the present chapter proposes the development of a service that will embody different functions –like content discovery, moderation and categorization- in a unique platform along with semantic annotation, which can contribute to content classification and verification. Journalists' education and subsequent engagement in design and evaluation of the service is considered to be of great importance for the success of the initiative.

REFERENCES

- Anderson, P. (2007). What is Web 2.0? Ideas, technologies and implications for education. *Technology & Standards Watch*, 1(1), 1–64.
- Beyers, H. (2004). Interactivity and online newspapers: A case study on discussion boards. *Convergence*, 10(4), 11–20. doi:10.1177/135485650401000403
- Boczkowski, J. B. (2005). *Digitizing the news: Innovation in online newspapers*. MIT Press.
- Bowman, S., & Willis, C. (2003). We media. How audiences are shaping the future of news and information. *The Media Center at the American Press Institute*. Retrieved August 8, 2015, from http://www.hypergene.net/wemedia/download/we_media.pdf
- Bradshaw, P., & Rohumaa, L. (2011). *The Online journalism handbook: Skills to survive and thrive in the digital age*. Longman.
- Bruns, A. (2005). *Gatewatching. Collaborative online news production*. Peter Lang.
- Bruns, A. (2008). The active audience: transforming journalism from gatekeeping to gatewatching. In C. Paterson & D. Domingo (Eds.), *Making online news: The ethnography of new media production* (pp. 171–184). Peter Lang.
- Deliyanni, E. (2012). Law and communication in blogs and new media. Athens: Nomiki Bibliothiki.
- Deuze, M. (2003). The Web and its journalisms: Considering the consequences of different types of newsmidia online. *New Media & Society*, 5(2), 203–230. doi:10.1177/1461444803005002004
- Deuze, M. (2006). Participation, remediation, bricolage: Considering principal components of a digital culture. *The Information Society: An International Journal*, 22(2), 63–75. doi:10.1080/01972240600567170
- Domingo, D., Quandt, T., Heinonen, A., Paulussen, S., Singer, J., & Vujnovic, M. (2008). Participatory journalism practices in the media and beyond: An international comparative study of initiatives in online newspapers. *Journalism Practice*, 2(3), 680–704. doi:10.1080/17512780802281065
- Franklin, B. (2014). The Future of Journalism. *Digital Journalism*, 2(3), 254–272. doi:10.1080/21670811.2014.930253
- Franquet, R., Villa, M. I., & Bergillos, I. (2011). Audience participation in online news websites: A comparative analysis. *Observatorio (OBS*)*, 5(3), 223–242.
- García-Avilés, J. A. (2010). “Citizen journalism” in European television websites: lights and shadows of user generated content. *Observatorio (OBS*)*, 4(4), 251–263.
- Gillmor, D. (2004). *We the Media. Grassroots journalism by the people, for the people*. O’Reilly.
- Hess, K. (2013). Breaking boundaries. Recasting the “local” newspaper as “geo-social” news in a digital landscape. *Digital Journalism*, 1(1), 48–63. doi:10.1080/21670811.2012.714933
- Hirst, M. (2011). *News 2.0. Can journalism survive the Internet?* Allen & Unwin.
- Johnson, B. G., & Dade, C. (2018). *Local broadcast journalism, user-generated content and boundary work*. Media Practice and Education., doi:10.1080/25741136.2018.1464742

Johnston, L. (2016). Social news=Journalism evolution? *Digital Journalism*, 4(7), 899–909. doi:10.1080/21670811.2016.1168709

Katsaounidou, A., & Dimoulas, C. (2018). Integrating content authentication support in media services. In M. Khosrow-Pour, D.B.A. (Ed.), *Encyclopedia of Information Science and Technology*, Fourth Edition (pp. 2908-2919). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-2255-3.ch254

Katsaounidou, A., Dimoulas, C., & Veglis, A. (2018). *Cross-media authentication and verification: Emerging research and opportunities*. IGI Global. doi:10.4018/978-1-5225-5592-6

Kim, Y., & Lowrey, W. (2015). Who are citizen journalists in the social media environment? *Digital Journalism*, 3(2), 298–314. doi:10.1080/21670811.2014.930245

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. In E. Dykstra-Erickson, & M. Tscheligi (Eds.), *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems* (pp. 543-550). Vienna, Austria: ACM. 10.1145/985692.985761

Lewis, S. C., Kaufhold, K., & Lasorsa, D. L. (2010). Thinking about citizen journalism: The philosophical and practical challenges of user-generated content for community newspapers. *Journalism Practice*, 4(2), 163–179. doi:10.1080/14616700903156919

Manosevitch, I. (2011). User generated content in the Israeli online journalism landscape. *Israel Affairs*, 17(3), 422–444. doi:10.1080/13537121.2011.584670

Mast, J., & Hanegreefs, S. (2015). When news media turn to citizen-generated images of war. *Digital Journalism*, 3(4), 594–614. doi:10.1080/21670811.2015.1034527

Meraz, S. (2009). The many faced “You” of social media. In Z. Papacharissi (Ed.), *Journalism and citizenship. New agendas in communication* (pp. 123–147). Routledge.

Moore, J. E., & Hatcher, J. A. (2018). Disrupting traditional news routines through community engagement. *Journalism Studies*, 19(1), 1–16. doi:10.1080/1461670X.2017.1423238

Moretzsohn, S. (2006). “Citizen journalism” and the myth of redemptive technology. *Brazilian Journalism Research*, 2(2), 29–46. doi:10.25200/BJR.v2n2.2006.81

Newman, N. (2009). The rise of social media and its impact on mainstream journalism: A study of how newspapers and broadcasters in the UK and US are responding to a wave of participatory social media, and a historic shift in control towards individual consumers. *Reuters Institute for the Study of Journalism*. Retrieved April 4, 2018, from: <https://ora.ox.ac.uk/objects/uuid:a980df14-1b49-401b-a136-78d47ab76cdc>

Nip, J. Y. M. (2006). Exploring the second phase of public journalism. *Journalism Studies*, 7(2), 212–236. doi:10.1080/14616700500533528

Paulussen, S., Heinonen, A., Domingo, D., & Quandt, T. (2007). Doing it together: Citizen participation in the professional news making process. *Observatorio (OBS*) Journal*, 1(3), 131–154.

Paulussen, S., & Ugille, P. (2008). User generated content in the newsroom: Professional and organisational constraints on participatory journalism. *Westminster Papers in Communication and Culture*, 5(2), 24–41. doi:10.16997/wpcc.63

- Reader, B. (2018). Working with the 'gated': A case study of ABC Open's blend of reciprocal journalism and 'collegial gatekeeping'. *Journalism*, 00(0), 1–17. doi:10.1177/1464884918767595
- Saridou, T., Panagiotidis, K., Tsipas, N., & Veglis, A. (2018). *Semantic tools for participatory journalism*. Paper presented at the 4th International Conference on Communication and Management (ICCM 2018), Athens, Greece.
- Saridou, T., Spyridou, L. P., & Veglis, A. (2017). Churnalism on the rise? *Digital Journalism*, 5(8), 1006–1024. doi:10.1080/21670811.2017.1342209
- Saridou, T., & Veglis, A. (2016). Participatory journalism policies in newspapers' websites in Greece. *Journal of Greek Media and Culture*, 2(1), 85–101. doi:10.1386/jgmc.2.1.85_1
- Siapera, E. (2012). Forms of online journalism and politics. In E. Siapera & A. Veglis (Eds.), *The handbook of global online journalism* (pp. 155–175). Wiley Blackwell. doi:10.1002/9781118313978.ch9
- Siapera, E., & Veglis, A. (2012). Introduction: The evolution of online journalism. In E. Siapera & A. Veglis (Eds.), *The handbook of global online journalism* (pp. 1–17). Wiley-Blackwell. doi:10.1002/9781118313978.ch1
- Singer, J. B., Hermida, A., Domingo, D., Heinonen, A., Paulussen, S., Quandt, T., Reich, Z., & Vujnovic, M. (2011). *Participatory journalism. Guarding open gates at online newspapers*. Wiley-Blackwell. doi:10.1002/9781444340747
- Sinton, M. (2018). No longer one-to-many: How Web 2.0 interactivity is changing public service radio's relationship with its audience. *Journal of Radio & Audio Media*, 25(1), 62–76. doi:10.1080/19376529.2017.1370713
- Spyridou, L. P. (2018). Analyzing the active audience: Reluctant, reactive, fearful, or lazy? Forms and motives of participation in mainstream journalism. *Journalism*, 00(0), 1–21. doi:10.1177/1464884918784733
- Stovall, J. B. (2004). *Web journalism: Practice and promise of a new medium*. Pearson.
- Suau, J., & Masip, P. (2014). Exploring participatory journalism in Mediterranean countries. *Journalism Practice*, 8(6), 670–687. doi:10.1080/17512786.2013.865964
- Thurman, N. (2008). Forums for citizen journalists? Adoption of user generated content initiatives by online news media. *New Media & Society*, 10(1), 139–157. doi:10.1177/1461444807085325
- Tong, J. (2015). Chinese journalists' views of user-generated content producers and journalism: A case study of the boundary work of journalism. *Asian Journal of Communication*, 25(6), 600–616. doi:10.1080/01292986.2015.1019526
- Turner, G. (2010). *Ordinary people and the media: the demotic turn*. Sage.
- Valcke, P., & Lenaerts, M. (2010). Who's author, editor and publisher in user-generated content? Applying traditional media concepts to UGC providers. *International Review of Law, Computers & Technology*, 24(1), 119–131. doi:10.1080/13600861003644533
- Veglis, A. (2013). Education of journalists on ICTs issues and opportunities. *Journal of Applied Journalism & Media Studies*, 2(2), 265–279. doi:10.1386/ajms.2.2.265_1

Veglis, A. (2014). Moderation Techniques for Social Media Content. In G. Meiselwitz (Ed.), Lecture Notes in Computer Science: Vol. 8531. *Social Computing and Social Media. SCSM 2014* (pp. 137–148). Springer. doi:10.1007/978-3-319-07632-4_13

Veglis, A., & Bratsas, C. (2018). Data Journalism. In M. Khosrow-Pour, D.B.A. (Ed.), Encyclopedia of Information Science and Technology, Fourth Edition (pp. 1196–1205). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-2255-3.ch103

Ward, M. (2002). *Journalism online*. Focal Press.

Weaver, D. H., & Willnat, L. (2016). Changes in U.S. Journalism. *Journalism Practice*, 10(7), 844–855. doi:10.1080/17512786.2016.1171162

ADDITIONAL READING

Aitamurto, T. (2013). Balancing between open and closed. Co-creation in magazine journalism. *Digital Journalism*, 1(2), 229–251. doi:10.1080/21670811.2012.750150

Bro, P., & Wallberg, P. (2015). Gatekeeping in a digital era. *Journalism Practice*, 9(1), 92–105. doi:10.1080/17512786.2014.928468

Ekman, M., & Widholm, A. (2015). Politicians as Media Producers. *Journalism Practice*, 9(1), 78–91. doi:10.1080/17512786.2014.928467

Hermida, A., & Thurman, N. (2008). A clash of cultures. *Journalism Practice*, 2(3), 343–356. doi:10.1080/17512780802054538

Lewis, S. (2012). The tension between professional control and open participation. *Information Communication and Society*, 15(6), 836–866. doi:10.1080/1369118X.2012.674150

McCluskey, M., & Hmielowski, J. (2011). Opinion expression during social conflict: Comparing online reader comments and letters to the editor. *Journalism*, 13(3), 303–319. doi:10.1177/1464884911421696

Popek, E. (2011). *Understanding the world of user-generated content*. The Rosen Publishing Group.

Schifferes, S., Newman, N., Thurman, N., Corney, D., Göker, A., & Martin, C. (2014). Identifying and verifying news through social media. *Digital Journalism*, 2(3), 406–418. doi:10.1080/21670811.2014.892747

Siapera, E. (2012). *Understanding new media*. Sage.

Singer, J. B. (2013). User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society*, 16(1), 55–73. doi:10.1177/1461444813477833

KEY TERMS AND DEFINITIONS

Citizen Journalism: The form of journalism in which citizens hold the news production process without professional intervention.

Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA):

A test used to identify human Internet users in order to block computer-derived mass entries to a service.

Gatekeeping: The process of filtering information to decide whether it is suitable for becoming news and being disseminated.

Moderation: The process of checking UGC which is submitted to a media website.

Participatory Journalism: The form of journalism that takes place when content is co-created by journalists and users on an editorially controlled platform and when the content created by users elsewhere is then acquired and used by professional media outlets.

Registration: The process of submitting personal information in order to get access to participatory applications.

User-Generated Content (UGC): Online content in many types produced by users and made available for publication.

Section 9

Engineering Science and Technologies

Adapting a Requirements Engineering Process by Key Factors Estimation

9

Graciela Dora Susana Hadad <https://orcid.org/0000-0003-4909-9702>*Universidad Nacional del Oeste, Argentina & Universidad de Belgrano, Argentina***Jorge Horacio Doorn***Universidad Nacional de La Matanza, Argentina & Universidad Nacional de Tres de Febrero, Argentina***Viviana Alejandra Ledesma***Universidad Nacional de La Matanza, Argentina*

INTRODUCTION

The monolithic application of any process, disregarding the context conditions, may lead to unnecessary lack of effectiveness. By the contrary, the adaptation of any process to a particular situation is considered a good practice in many fields. Literature shows that this practice is quite common in Software Engineering processes, such as the methodologies Rational Method Composer (Haumer, 2005) and OPEN Process Framework (Firesmith & Henderson-Sellers, 2002). However, Requirements Engineering (RE) approaches are seldom tailored to context or project situations (Potts, 1995; Leite, Hadad, Doorn, & Kaplan, 2000; Leffingwell & Widrig, 2003; Seyff et al., 2009). Nevertheless, sometimes the elicitation activity, as part of an RE process, is performed taking into account some environmental characteristics, usually called situational factors, such as number of information sources, users geographical distribution, users time availability, users experience, among others (Maiden & Rugg, 1996; Hickey & Davis, 2003; Coulin, 2007; Carrizo, Dieste, & Juristo, 2008). Recently, some proposals have appeared to design an RE process for a specific project by selecting existent RE techniques (Lauesen, 2002; Lobo & Arthur, 2005; Alexander & Beus-Dukic, 2009).

There are activities of most of the requirements processes that are invariant regardless of situational factors, while others should be modified, removed or replaced. Not only activities may be adapted, models created in the process may be also suited for the situation (Galster, Weyns, Tofan, Michalik & Avgeriou, 2014). This means that these processes may be assembled like a flexible puzzle using interchangeable pieces depending on the situational factors identified.

Situational Method Engineering (SME) is advocated to build methods tailored to specific situations for the development of systems (Kumar & Welke, 1992). Following its principles, the adaptation of any software development process is based on indicators describing the situation (Khan, bin Mahrin & bt Chuprat, 2014). Part of the task is to compose such indicators based on observable factors, like degree of business processes reengineering, context complexity, developer expertise in the application domain, and project size, among others. Ideally, these situational factors should be considered before beginning the software process. However, there are factors not accurately known when initiating a software project, while other factors may change during the project. Hence, a dynamic view of the adaptation of a software

DOI: 10.4018/978-1-7998-3479-3.ch081

development process achieves a better performance of the process itself. Considering that defining requirements is the starting point of a software development, it should be necessary to pay more attention to factors influencing the RE process.

A frequent question of practitioners is related with the need of performing all the process steps to develop the software requirements. *Is it possible to shorten the road or to follow a different one?* Under some circumstances, there is an opportunity to reduce the RE process by deleting or simplifying activities; and sometimes different paths may be followed by choosing other techniques or even extending some activities. Project managers should make decisions depending mainly on the reality he or she is facing.

Therefore, the rational and some practice on the tailoring of an RE process according to a particular set of situational factors is presented in this chapter. Recommendations about the estimation of these factors are exposed as an enhanced solution. Additionally, some lessons learned, and future works are reported.

BACKGROUND

To follow a process guiding the development of an engineering project or system is valuable enough since it means that performing a predictable set of activities using techniques helps to get the outcome within a controllable quality.

Hence, the way the work is performed does not depend on individual criteria, allowing repeatability of costs, times and quality, and promoting the accumulation of knowledge about the process. The first activity of a process to develop a product consists in defining, as precisely as desirable, the expected outcome. When the product is a software system, this initial activity is an RE process, whose outcome is a consistent set of requirements. The RE process is particularly different from other activities of the software development process since it is the one that most interacts with people and their environment, while other activities are mainly carried out within the development team (Carrizo, 2009). Besides, project decisions impose constraints, tools and methods to carry out those activities. Therefore, if the requirements process takes into account the particularities surrounding the application context and the project itself, then it will probably result both in a better set of requirements and in a more efficient process.

Furthermore, a requirements process needs appropriate and continuous communication to gain as much customers and users compromise as possible. A better communication is achieved when all stakeholders use the same language. In RE, a proven way to accomplish this is by using the vocabulary of the application context (Leite, Doorn, Kaplan, Hadad, & Rida, 2004). Communication occurs when stakeholders orally interact, also when reports, documents and models are exhibited to customers. In this sense, natural language (NL) models, such as glossaries, use cases and scenarios, stimulate stakeholders' communication (Leite et al., 2004), and they are the most frequently used in RE (Kaindl, 2000; Leffingwell & Widrig, 2003; Seyff et al., 2009; Jacobson, Spence, & Bittner, 2011; Antonelli, Rossi, Leite, & Oliveros, 2012).

Software development processes put into practice in real projects are often forced to be adjusted due to contingent circumstances, sometimes in a poorly controlled fashion while the projects are ongoing. Hence, the situations within their application context and projects should be observed as the processes go forward for better tailoring to such evolution. Possible adjustments to the processes can be known in advance based on certain characteristics, though they may change dynamically, i.e., settings are pre-planned but only implemented when an aspect of the situation changes (Rolland, 2008). In this regard, the process may be defined as a set of blocks, having process blocks common to all situations and variant process blocks according to situational factors. Thus, the process is made up by assembling blocks for

the particular situation (Henderson-Sellers & Ralyté, 2010). Method Engineering has emerged to tackle this case, promoting the design, construction and adaptation of methods, techniques and tools in order to develop information systems (Brinkkemper, 1996). This discipline considers not only the creation of process blocks but also product blocks, and even blocks that assemble both process and product (Rolland, 2008; Henderson-Sellers & Ralyté, 2010; Ralyté, 2013). Rolland (2008) not only has pointed out that the software development should begin with a definition phase of the method to be used, but also she proposes that the process must be re-adjusted throughout its life cycle, an aspect not treated by Method Engineering. SME, as a sub-area of Method Engineering, focuses on building methods for software development, tailored to specific situations (Kumar & Welke, 1992). Thus, SME studies the factors affecting the software project and the application context (Bucher, Klesse, Kurpjuweit, & Winter, 2007). Its principles have been applied in RE to define requirements processes that are adaptable to particular circumstances by using existent modular components (Firesmith, 2004; Jafarinezhad & Ramsin, 2012; Bakhat, Sarwar, Motla, & Akhtar, 2015). These works can be grouped as follows:

- Group A:** Works oriented to define an RE process using frameworks. A knowledge base on RE activities and techniques is made available, serving as a guide to obtain an RE process according to the project's needs (Olsson, Doerr, Koenig & Ehresmann, 2005; Zowghi, Firesmith, & Henderson-Sellers, 2005; Jiang & Eberlein, 2008).
- Group B:** Approaches to select elicitation techniques that best suit to the situational context. These works only focus on the elicitation activity in the RE process (Hickey & Davis, 2004; Carrizo, Dieste, & Juristo, 2008; Al-Zawahreh & Almakadmeh, 2015).
- Group C:** Models that allow the RE process to be adapted when it is applied to a specific class of project. These are proposals of situational adaptation for projects of a certain nature or scope, such as global software development or agile methods (Bakhat, Sarwar, Motla, & Akhta, 2015; Abdullah & Khan, 2015; Kabaale & Mayoka Kituyi, 2015; Khan, bin Mahrin, & Mali, 2016).
- Group D:** Guides that facilitate the creation of an RE process adapted to the context. These works provide mechanisms for the creation of an RE process adjusted to the situational characteristics of the environment, within projects of any nature (Lobo & Arthur, 2005; Alexander & Beus-Dukic, 2009; Coulin, Zowghi, & Sahraoui, 2006; Jafarinezhad, & Ramsin, 2012), in contrast to works of Group C.

These works use notions of SME in different degrees of adherence to its principles, and they propose some factors to be evaluated in order to define the situation. Particularly, approaches in Groups A and D use process components to conform the new RE process. However, only proposals in Group D combine the concepts of SME with some principles of process variability.

The work presented by Hickey & Davis (2004), within Group B, poses a partially dynamic approach, an aspect that in general terms is not taken into account in other proposals. The authors emphasize that knowledge about the problem domain and the project domain is acquired as the RE process progresses, and thus, more appropriate elicitation techniques can be selected in each iteration. In addition, they mention that the situational characteristics of the problem context are usually static, while those related to the project may change, although they do not present evidence on the matter. In this sense, reminding what Rolland (2008) points out, contingent factors change during project life cycle, leading to a dynamic adaptation of the process.

Most of the studies related to the SME do not address two essential aspects: the incorrect adaptation of the process, and the probability of evolution or incorrect estimation of the factors that guide the adaptation. The first issue can occur under three variants: i) the inappropriate omission of doing an activity

that could be necessary, ii) the incorrect replacement of an activity by another, or iii) the unnecessary insertion of an activity. It is clear that the process adaptation is strongly dependent on the factors estimation, so a study of the evolution and estimation of situational factors can reveal issues not so easily visible, which can lead to an unsuccessful adaptation.

The work presented in this chapter is part of Group D. It aims to establish clear mechanisms to adapt an RE process according to particular characteristics of the application context and the project context, and to context evolution, considering that evolution may come not only from a real change of the situation but also from a better understanding of that situation.

Some of the adaptation factors are related to development time constraints. Naturally, they lead to faster ways of developing software; these are steps toward agile methods (Pineiro, 2002). A quality factor may depend on a time factor, or the other way around. However, a balance equation for both factors may be frequently achieved (Kohler & Paech, 2002; Pineiro, 2002). This trade-off is more likely to be better faced by an RE process with the necessary tailoring (Kohler & Paech, 2002). Extremely, when time factor is a major concern, agile methods could be the best way. Though, agile practices are frequently assumed as opposite to RE practices (Cockburn, 2002; Beck, 2004), Kovitz (2002) remarks that both, a phased development and an agile one, do RE but in different styles, since requirements are always present. This vision is also shared by (Sillitti & Succi, 2005; Kohler & Paech, 2002). In this sense, an agile approach dealing with requirements may be also seen as a sort of dynamic adaptation of an RE process. As Cockburn (2002) said, making the decision on the right level of agility is the best way to succeed on a project. RE practices do not deny agility, considering that this property must be combined with quality and attending volatility without constant software fix (Pineiro, 2002).

AN ADAPTABLE REQUIREMENTS ENGINEERING PROCESS

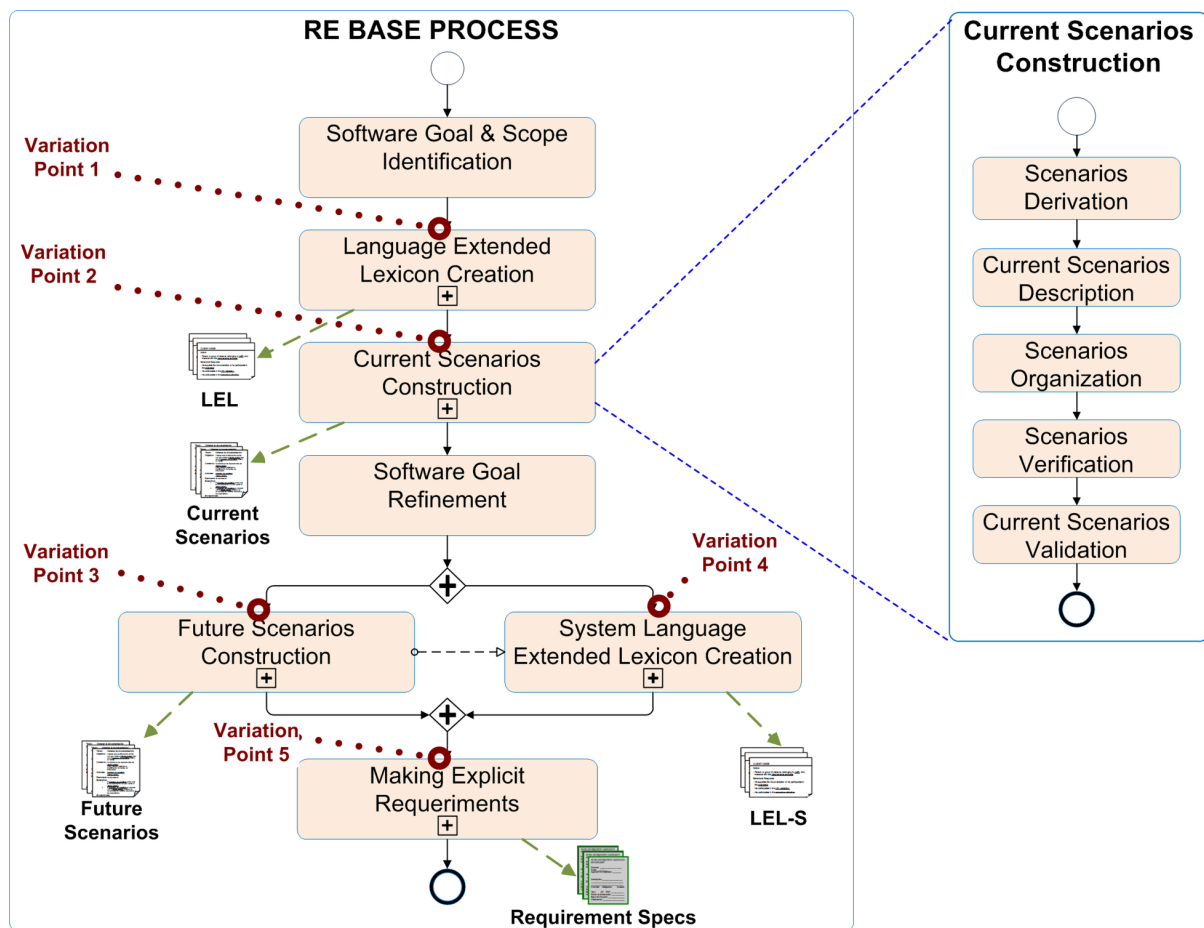
Tailoring a Requirements Base Process

RE process based on NL models strongly contributes to stakeholders' commitment to the project. As a standard or base process, it involves the following stages:

- **Identifying the general software goal and scope.**
- **Understanding the vocabulary used in the application context**, supported by the Language Extended Lexicon model (Hadad, Doorn, & Kaplan, 2009).
- **Understanding the application context**, supported by a set of current scenarios that represent the situations observed in the application context (Leite et al., 2000).
- **Refining the general software goal**, by decomposing it into sub-goals.
- **Defining the software context**, by producing a set of future scenarios that represent situations envisioned in a future application context where the software system will operate (Leite et al., 2004).
- **Defining the vocabulary used in software descriptions**, by producing a Language Extended Lexicon of the System (LEL-S) created from the previous one, adding terms used in the descriptions of future scenarios.
- **Making explicit the requirements**, by producing consistent requirement specs, after extracting requirements from future scenarios. Eventually, a Software Requirements Specification (SRS) document may be produced.

This RE process has been initially developed, tested and used in a monolithic fashion, as depicted in Figure 1(a), ignoring all special circumstances. Actually, this implies a somehow poorly defined standard context. It should be noticed that the base process in Figure 1(a) is decomposed in sub-processes, where most of them are composed of another level of processes, as shown in Figure 1(b). Recycles in the two diagrams of Figure 1 were removed to emphasize the variation points, which are introduced here only for the proposed adaptation solution. These recycles imply correcting or updating models because of a better understanding of the problem and of the evolution of the application context.

Figure 1. Requirements Engineering Base Process, expanding one main sub-process



Though Figure 1(a) shows a monolithic process as part of a waterfall model, the strategy can be adapted to be used within different software process models, being this a project factor by itself (see Table 1). For instance, when following an iterative or incremental process model, all sub-processes (except the first one) are done in iterations along with a requirements change management activity throughout the software development life cycle. In case of developing an incremental RE process, the increments may involve the enhancement of previous functionalities and properties, or the addition of new ones; this is treated by a specific factor, which may take different values in each cycle. Regarding agile practices, additionally to time constraints, another leading factor is volatility (Cockburn, 2002), which may be well

coped with a change management activity. However, it may not be the best manner when volatility is a core characteristic of the context. Therefore, an agile method might be a better proposal to satisfy customers.

This RE base process may be adapted at the variation points shown in Figure 1(a), depending on a combination of situational factors. Project managers should do the adaptation to improve the RE process and, thereby, the requirements product. Situational factors impacting on RE (see Table 1) have been selected based on experience in RE practice and literature proposals (Maiden & Rugg, 1996; Hickey & Davis, 2003; Carrizo et al., 2008; Ebling, Audy, & Prikladnicki, 2009; Jafarinezhad & Ramsin, 2012; Mighetti & Hadad, 2016). These factors can be characterized by the following attributes:

- **Value:** Admissible values depend on each factor
- **Confidence on the value:** Sure, Reasonably Sure or Doubtful
- **Origin:** Application Context or Project
- **Evolution type:** Invariant or Contingent
- **Variation points:** Where the factor impacts on

In order to adapt the RE process, project managers (or method engineers) assign values to factors, qualifying them with their degree of confidence on such values. It usually happens that at the very beginning some factors may be not precisely known, and right values are known after gaining a better understanding of the situation. Factors are classified according to their origin into those related to the specific application context, and those related to the specific software project. From an evolution dimension, factors are classified into: *Invariant*, being those that do not change during the RE process; and *Contingent*, being those that may naturally evolve due to changes in the application context and/or in the project context.

Interaction among factors should be taking into account when tailoring the RE process. *Neutral interaction* implies that two factors are fully independent, having no influence between them. *Overriding interaction* indicates that a factor is discarded when another factor takes a specific value. *Limiting interaction* points out that a factor may reduce the range of acceptable values of another factor. *Incompatible interaction* implies that the values of two or more factors cannot be accepted simultaneously at the same project; thereby, the value of at least one factor has to be changed through stakeholders' negotiation to solve the inconsistency. Examples of these interactions are:

- **Overriding interaction:** The factor Business Process Reengineering is not taken into account if the factor Novelty of the Context takes the value *Yes*.
- **Limiting interaction:** The factor Software Process Model cannot take the value *Waterfall* if factor Context Volatility is *Very High*.
- **Incompatible interaction:** The factor Software Reliability is *Very High* while the factor Project Time and Resources Constraints is *Very Low*.

Each sub-process is defined by a sequence of process blocks, some of which include product models. At each variation point, a base process block may: i) stay as it is; ii) be deleted; iii) be replaced by another process block; or iv) be replaced by a process block with a partial internal variation based on parameters. Figure 2, exploiting sub-process of Figure 1(b), shows how the instance of a process should be constructed according to specific situational factors, using a repository containing all possible atomic blocks. This Figure exemplifies one process block (*Deriving Current Scenarios*) that may be replaced by one of three different atomic processes and, by the way, all process blocks are parameterized.

Table 1. Situational Factors impacting on RE process. Range Very High to Very Low includes five possible values

Context Factors	Acceptable Values	Type of Evolution
Context Complexity	Very High to Very Low	Invariant
Target Customer	Taylor-made, Market-driven	Invariant
Novelty of the Context	Yes, No	Invariant
Business Process Reengineering	Very High to Very Low	Contingent
Context Volatility	Very High to Very Low	Invariant
Volatility of Users Needs	Very High to Very Low	Invariant
Inconsistencies in the Context	Very High to Very Low	Invariant
Conflict among Users Interests	Very High to Very Low	Contingent
Users Rotation	Very High to Very Low	Invariant
Users Location	Co-located, Distributed	Invariant
Project Factors	Acceptable Values	Type of Evolution
Familiarity with the Domain	Very High to Very Low	Invariant
Project Size	Very High to Very Low	Contingent
Developers Rotation	Very High to Very Low	Invariant
Software Reliability	Very High to Very Low	Invariant
Reuse of Existing Requirements Artifacts	LEL, Current Scenarios, Future Scenarios, LEL-S, Requirements, No	Invariant
Demand to Create Requirements Artifacts for Reuse	LEL, Current Scenarios, LEL-S, Requirements, No	Contingent
Demand of Requirements Pre-Traceability	Total, Partial, No	Invariant
Demand of Requirements Post-Traceability	Yes, No	Invariant
Granularity of Requirements Traceability	Individual, Group	Invariant
Demand to Produce an SRS	Yes, No	Invariant
Project Time and Resources Constraints	Very High to Very Low	Contingent
Software Process Model	Waterfall, Incremental, Iterative, Formal Transformation, Other	Invariant
Type of Software Increment	New Functionality, Enhanced Functionality	Contingent
Experience in RE Process	Very High to Very Low	Invariant
Cultural and Language Barriers	Very High to Very Low	Invariant
Time Differences among Stakeholders	Very High to Very Low	Invariant

There are common elements among the various parts of the RE process. For example, Figure 2 shows three process blocks *Scenarios Derivation*, *Scenarios Organization* and *Scenarios Verification* which are shared with another main sub-process *Future Scenarios Construction*. There are four types of atomic process blocks. An *own block* is that not shared by any other main sub-process and is independent of situational factors. A *shared block* is common to different sub-processes of the RE process and may depend on factors. A *variant block* belongs to a specific sub-process and depends on situational factors. A *null block* is an auxiliary block used when an entire sub-process or an atomic process must be deleted or skipped from the basic RE process. Table 2 shows operations that may be applied to each type of process block.

Figure 2. Defining the Current Scenarios Construction Sub-Process

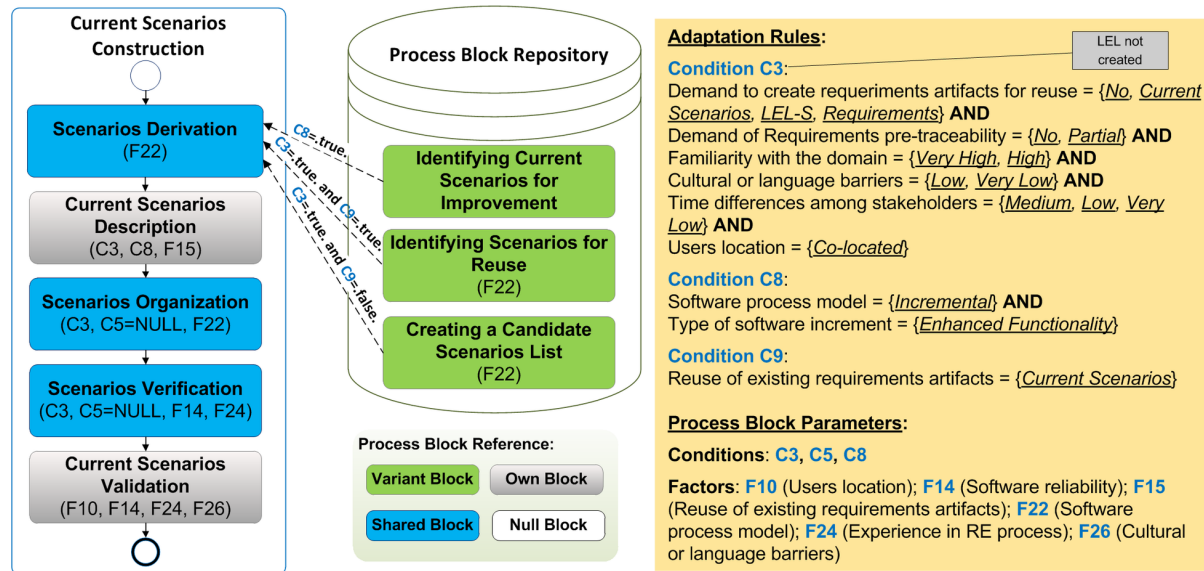


Table 2. Valid Operations on Basic Process Blocks

BLOCK TYPE	OPERATION			
	Stay	Delete	Replace	Parameterize
Own	✓	✓		✓
Shared	✓	✓		✓
Variant		✓	✓	

Some process blocks include product models, which contain minor variations of the NL models produced by the RE process. For example, if the lexicon model is not created, then the scenario model does not include hyperlinks to lexicon symbols.

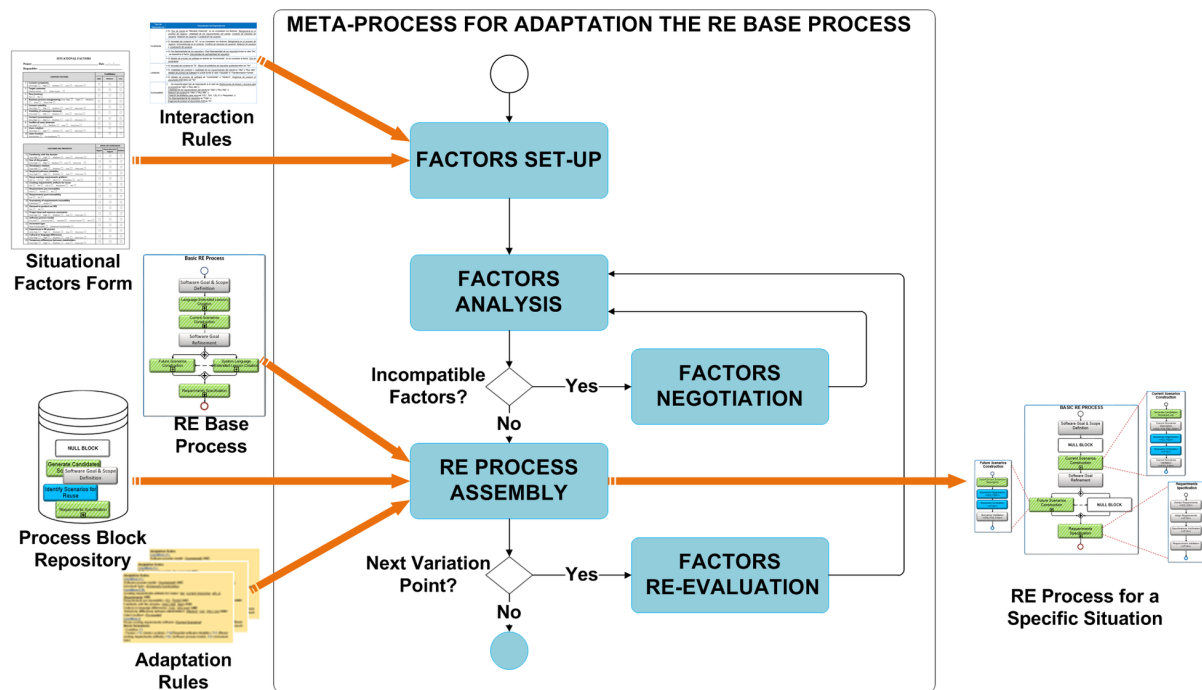
Meta-Process for a Dynamic Adaptation of a Requirements Process

The adaptation of the RE base process involves a mechanism of operations on process blocks at every variation point, using a specific combination of situational factors, following SME principles. The variation points are milestones used to review the RE process and to indicate possible re-designs of remaining RE activities. The way of constructing the RE process for a specific situation includes the following activities (see Figure 3):

1. **Factors Set-up:** The situational factors are evaluated, assigning a degree of confidence to each value, taking into account the interaction among factors. A simple form is used to collect the factors value. This form eases its filling with valid values by means of rules for factors interaction.
2. **Factors Analysis:** The factors setting is analyzed considering mainly interactions and factors with low confidence. Project managers should take a position on every factor with low confidence: optimistic, conservative or balanced position. Under an optimistic position, factors are assigned an

- extreme value of its valid range depending on the meaning of the factor, while under a conservative position factors take the opposite extreme value. A balanced position involves choosing intermediate values. This re-allocation of values is done to allow assembling the process, since the original values are preserved. When incompatible interactions are observed, a negotiation must take place.
3. **Factors Negotiation:** The purpose of this activity is to reach an agreement among stakeholders by modifying the causes that bring out the incompatibility. Then a re-assignment of values to one or more of the involved factors is done. After negotiation, a partial analysis of these factors takes place.
 4. **RE Process Assembly:** The RE process is assembled departing from the base process, considering a subset of situational factors at each variation point. A set of rules based on a specific combination of those involved factors defines which operation is applied over each base process block (see an example at Figure 2, right part). Thus, this assembly activity consists in maintaining, deleting, replacing, or parameterizing process blocks.
 5. **Factors Re-Evaluation:** At every variation point, those factors involved in remaining variation points must be re-evaluated. Low-confidence factors must be watched, since its level of confidence is supposed to increase due to a better perception of the situation. Contingent factors must be also watched to establish if a change in the situation has occurred. After this activity, a new factors analysis is needed in order to ensure the absence of inconsistencies among dependent factors.

Figure 3. Process for Dynamically Tailoring an RE Process, based on situational factors



Issues on Situational Factors Evaluation

The evolution of 21 situational factors has been studied during the developing of the RE Process in 35 industry projects. This study has allowed corroborating the evolution of the situations and, addition-

ally, to identify factors to which more attention must be paid for a suitable dynamic adaptation of the requirements process.

This study had confirmed some underlined initial hypothesis, such as:

- The confidence increased along the Requirements Engineering process: the initial percentage of doubtful factors was 12.8 and near to the ending, this figure was 4.7.
- The confidence on the context factors is higher than the confidence on the project factors: the initial percentages of doubtful factors were 7.2 and 17.0 respectively, while near o the ending they were 3.0 and 5.9.
- High level of variation along the project occurred in some presumed contingent factors. Most relevant variations were on factors: Business Process Reengineering with 54% of changes and Project Size with a variation of 46%.

However, some unexpected observations also occurred:

- There are still doubts on the value of some factors when the RE process was almost finishing, with a notorious 12% of Demand of Requirements Post-Traceability, 10% on Volatility of Users Needs and 9% on Project Size.
- The overall percentage of factors with relevant change was of 33.2, which is higher than expected; 86,5% of them were obtained with high degree of confidence. These does not discard estimation errors.
- High level of variation along the project occurred in some presumed invariant factors. Most relevant were Software Reliability with 53% of changes with an 87% of no doubtful estimation, and Demand of Requirments Pre-Traceability with 44% of changes in a context of 20% of doubts in the estimation.

SOLUTIONS AND RECOMMENDATIONS

Since selecting the appropriate RE process for a specific situation is crucial, project managers should define as precisely as possible the situational factors in order to construct a suitable RE process. However, they should be aware of the potential evolution of the situation and, thereby, review factors and adjust the RE process at pre-defined milestones.

Evolution of the situation is a key matter that must not be underestimated. Factors may naturally change due to context evolution, or some may require an update due to unprecise prior estimations. Therefore, the quality of estimating situational factors is critical for a satisfactory RE process adaptation.

The analysis of the results of the factors evaluation study lead to recognize that a good quality estimation may be not so easy to do at early stages

Factors estimation requires new evaluations while performing the RE process adaptation. This activity must be mainly focus on those most unpredictable factors (contingent factors), and on those whose values are more difficult to perceive. Furthermore, when reevaluation takes place, it should to be taken into consideration that changes in some factors may impact changes over others, and these changes should not be incompatible with the part of the RE process done before.

The factors evolution study has revealed that uncertain or incorrect factors estimations are mainly due to lack of or poor guides for developing the RE activities and, somehow, due to a poor factors definition.

FUTURE RESEARCH DIRECTIONS

Feasible variants of the process were identified due to the experience achieved after putting into practice the RE base process in many software projects in industry along more than two decades.

Situational factors were exhaustively studied in literature, although this set is likely to be extended. A first extension was made since the previous work published in (Hadad, Doorn & Ledesma, 2018). Factors affecting subsequent phases of the software development process will be studied to determine if they may influence backward on the RE process.

The decomposition of the RE process in base process blocks, along with the set of situational factors, has allowed identifying process commonality and process variability, and hence defining the different types of process blocks needed. It is planned to evaluate the convenience of using mixed blocks, or separated process blocks from model blocks. It is also considered that more cases should be done to confirm the different branches of the adaptable RE process.

New studies focusing on the weakness already detected on factors estimation will be carried out.

CONCLUSION

The proposal presented in this chapter deals with variability in an RE process, driven by a dynamic adaptation based on prevailing situational factors.

Most problems have distinctive features that must be taken into account to carry out a successful requirements process. Project managers should tailor the requirements process by choosing the techniques that are most suitable for specific situations. When guidelines on how to adapt the requirements process are provided, managers are more willingly to do it.

The meta-process for constructing an RE process involves a set of factors that typifies situations considering both the application context and the software project context. This adaptation process establishes when the RE process should be reviewed and re-tailored based on rules and operations at each variation point. Since it is frequently observed the natural and contingent evolution of situations, it is important to have the possibility of redefining the remaining activities of the RE process once the project is in course. Focus should be put on the possibility of an incorrect evaluation of certain factors, being aware of the necessity of re-adaptations to avoid a bias towards an unsuccessful process.

The adaptation of the requirements process for several real cases has contributed to its acceptance in host organizations. It is considered that the dynamic adaptation process could be implemented with minor adjustments in other RE strategies.

REFERENCES

- Abdullah, A., & Khan, H. (2015). FreGsd: A Framework for Global Software Requirement Engineering. *JSW*, 10(10), 1189–1198. doi:10.17706/jsw.10.10.1189-1198
- Al-Zawahreh, H., & Almakadmeh, K. (2015). Procedural Model of Requirements Elicitation Techniques. In *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication* (article 65). ACM. doi:10.1145/2816839.2816902

Alexander, I. F., & Beus-Dukic, L. (2009). *Discovering Requirements: How to Specify Products and Services*. John Wiley & Sons.

Antonelli, L., Rossi, G., Leite, J. C. S. P., & Oliveros, A. (2012). Deriving requirements specifications from the application domain language captured by Language Extended Lexicon. In *Proceedings of XV Workshop on Requirements Engineering*. Buenos Aires, Argentina: Universidad Nacional de La Matanza.

Bakhat, K. A., Sarwar, A. A., Motla, Y. H. B., & Akhtar, M. C. (2015). A Situational Requirement Engineering Model for an Agile Process. *Bahria University Journal of Information & Communication Technology*, 8(1), 21–26.

Beck, K. (2004). *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional.

Brinkkemper, S. (1996). Method Engineering: Engineering of Information Systems Development Methods and Tools. *Information and Software Technology*, 38(4), 275–280. doi:10.1016/0950-5849(95)01059-9

Bucher, T., Klesse, M., Kurpjuweit, S., & Winter, R. (2007). Situational Method Engineering. In *Situational method engineering: fundamentals and experiences* (pp. 33-48). Springer US. doi:10.1007/978-0-387-73947-2_5

Carrizo, D. (2009). *Marco para la selección de técnicas para educación de requisitos* (Doctoral Thesis). Universidad Politécnica de Madrid, Madrid, Spain.

Carrizo, D., Dieste, O., & Juristo, N. (2008). Study of elicitation techniques adequacy. In *Proceedings XI Workshop on Requirements Engineering* (pp. 104-114). Barcelona, Spain: Universitat Politècnica de Catalunya.

Cockburn, A. (2002). *Agile Software Development*. Addison-Wesley.

Coulin, C., Zowghi, D., & Sahraoui, A. (2006). A Situational Method Engineering Approach to Requirements Elicitation Workshops in the Software Development Process. *Software Process Improvement and Practice*, 11(5), 451–464. doi:10.1002/pip.288

Coulin, C. R. (2007). *A Situational Approach and Intelligent Tool for Collaborative Requirements Elicitation* (Doctoral Thesis). University of Technology, Sydney, Australia.

Ebling, T., Audy, J. L. N., & Prikladnicki, R. (2009). Towards a Requirements Reuse Method Using Product Line in Distributed Environments. In *Proceedings of 12th Workshop on Requirements Engineering* (pp. 91-102). Valparaíso, Chile: Universidad Técnica Federico Santa María.

Firesmith, D. (2004). Creating a project-specific requirements engineering process. *Journal of Object Technology*, 3(5), 31–44. doi:10.5381/jot.2004.3.5.c4

Firesmith, D. G., & Henderson, B. (2002). *The OPEN Process Framework: An Introduction*. Addison-Wesley.

Galster, M., Weyns, D., Tofan, D., Michalik, B., & Avgeriou, P. (2014). Variability in Software Systems - A Systematic Literature Review. *IEEE Transactions on Software Engineering*, 40(3), 282–306. doi:10.1109/TSE.2013.56

Hadad, G. D. S., Doorn, J. H., & Kaplan, G. N. (2009). Creating Software System Context Glossaries. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (2nd ed., pp. 789–794). IGI Global. doi:10.4018/978-1-60566-026-4.ch128

Hadad, G. D. S., Doorn, J. H., & Ledesma, V. A. (2018). Dynamic Situational Adaptation of a Requirements Engineering Process. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (4th ed., pp. 7422–7434). IGI Global, Information Science Reference. doi:10.4018/978-1-5225-2255-3.ch646

Haumer, P. (2005). *IBM Rational Method Composer: Part 1: Key concepts*. Retrieved June 28, 2019, from <https://www.ibm.com/developerworks/rational/library/dec05/haumer/haumer-pdf.pdf>

Henderson-Sellers, B., & Ralyté, J. (2010). Situational Method Engineering: State-of-the-Art Review. *J. UCS*, 16(3), 424–478.

Hickey, A. & Davis, A. (2004). A Unified Model of Requirements Elicitation. *Journal of Management Information Systems*, 20(4), 65–84. doi:10.1080/07421222.2004.11045786

Hickey, A. M., & Davis, A. M. (2003, September). Elicitation technique selection: how do experts do it? In *Proceedings of 11th IEEE International Requirements Engineering Conference* (pp. 169–178). IEEE. 10.1109/ICRE.2003.1232748

Jacobson, I., Spence, I., & Bittner, K. (2011). USE-CASE 2.0: The Guide to Succeeding with Use Cases. *Ivar Jacobson International*. Retrieved June 28, 2019, from https://www.ivarjacobson.com/sites/default/files/field_iji_file/article/use-case_2_0_jan11.pdf

Jafarinezhad, O., & Ramsin, R. (2012, March). Towards a process factory for developing situational requirements engineering processes. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 1089–1090). ACM. 10.1145/2245276.2231946

Jiang, L., & Eberlein, A. (2008). A Framework for Requirements Engineering Process Development (FRERE). In *Proceedings of 19th Australian Conference on Software Engineering* (pp. 507–516). Perth, Australia. 10.1109/ASWEC.2008.4483240

Kabaale, E., & Mayoka Kituyi, G. (2015). A Theoretical Framework for Requirements Engineering and Process Improvement in Small and Medium Software Companies. *Business Process Management Journal*, 21(1), 80–99. doi:10.1108/BPMJ-01-2014-0002

Kaindl, H. (2000). A design process based on a model combining scenarios with goals and functions. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans*, 30(5), 537–551. doi:10.1109/3468.867861

Khan, H., bin Mahrin, M., & Mali, M. (2016). Situational Requirement Engineering Framework for Global Software Development: Formulation and Design. *Bahria University Journal of Information & Communication Technologies*, 9(1), 74–84. doi:10.1109/I4CT.2014.6914179

Khan, H.H., bin Mahrin, M.N., & bt Chuprat, S. (2014). Factors for Tailoring Requirements Engineering Process: A Review. *International Journal of Software Engineering and Technology*, 1(1), 7–18.

Kohler, K., & Paech, B. (2002). Requirement Documents that Win the Race: Not Overweight or Emaciated but Powerful and in Shape. *International Workshop on Time-Constrained Requirements Engineering (TCRE'02)*.

Kovitz, B. L. (2002). Hidden Skills that Support Phased and Agile Requirements Engineering. *International Workshop on Time-Constrained Requirements Engineering (TCRE'02)*.

Kumar, K., & Welke, R. J. (1992). Methodology Engineering: a proposal for situation-specific methodology construction. In W. W. Cotterman & J. A. Senn (Eds.), *Challenges and strategies for research in systems development* (pp. 257–269). John Wiley & Sons.

Lauesen, S. (2002). *Software Requirements: Styles and Techniques*. Addison-Wesley.

Leffingwell, D., & Widrig, D. (2003). *Managing Software Requirements: a unified approach*. Addison-Wesley Professional.

Leite, J. C. S. P., Doorn, J. H., Kaplan, G. N., Hadad, G. D., & Ridao, M. N. (2004). Defining System Context using Scenarios. In J. C. S. P. Leite & J. H. Doorn (Eds.), *Perspectives on Software Requirements* (pp. 169–199). Springer. doi:10.1007/978-1-4615-0465-8_8

Leite, J. C. S. P., Hadad, G. D., Doorn, J. H., & Kaplan, G. N. (2000). A scenario construction process. *Requirements Engineering*, 5(1), 38–61. doi:10.1007/PL00010342

Lobo, L. O., & Arthur, J. D. (2005, April). An objectives-driven process for selecting methods to support requirements engineering activities. In *Proceedings of 29th Annual IEEE/NASA Software Engineering Workshop* (pp. 118–130). IEEE. 10.1109/SEW.2005.18

Maiden, N. A. M., & Rugg, G. (1996). ACRE: Selecting methods for requirements acquisition. *Software Engineering Journal*, 11(3), 183–192. doi:10.1049ej.1996.0024

Mighetti, J. P., & Hadad, G. D. S. (2016). A Requirements Engineering Process Adapted to Global Software Development. *CLEI Electronic Journal*, 19(3), paper 7. doi:10.19153/cleiej.19.3.7

Olsson, T., Doerr, J., Koenig, T., & Ehresmann, M. (2005). A Flexible and Pragmatic Requirements Engineering Framework for SME. In *Proceedings of 1st International Workshop on Situational Requirements Engineering Processes (SREP'05)* (pp. 1–12). Paris, France: Academic Press.

Pinheiro, F. (2002). Requirements Honesty. *International Workshop on Time-Constrained Requirements Engineering (TCRE'02)*.

Potts, C. (1995, August). Using schematic scenarios to understand user needs. In *Proceedings of the 1st conference on Designing interactive systems: processes, practices, methods, & techniques* (pp. 247–256). ACM. 10.1145/225434.225462

Ralyté, J. (2013). Situational Method Engineering in Practice: A Case Study in a Small Enterprise. *Proceedings of CAiSE'13 Forum at the 25th International Conference on Advanced Information Systems Engineering*, 17–24.

Rolland, C. (2008). Method engineering: towards methods as services. In *Making Globally Distributed Software Development a Success Story* (pp. 10–11). Springer Berlin Heidelberg. doi:10.1007/978-3-540-79588-9_2

Seyff, N., Maiden, N., Karlsen, K., Lockerbie, J., Grünbacher, P., Graf, F., & Ncube, C. (2009). Exploring how to use scenarios to discover requirements. *Requirements Engineering*, 14(2), 91–111. doi:10.1007/00766-009-0077-9

Sillitti, A., & Succi, G. (2005). Requirements Engineering for Agile Methods. In A. Aurum & C. Wohlin (Eds.), *Engineering and Managing Software Requirements* (pp. 309–326). Springer-Verlag. doi:10.1007/3-540-28244-0_14

Zowghi, D., Firesmith, D., & Henderson-Sellers, B. (2005). Using the OPEN Process Framework to Produce a Situation-Specific Requirements Engineering Method. In *Proceedings of 1st International Workshop on Situational Requirements Engineering Processes* (pp.59-74). Paris, France: Academic Press.

ADDITIONAL READING

Browne, G. J., & Ramesh, V. (2002). Improving information requirements determination: A cognitive perspective. *Information & Management*, 39(8), 625–645. doi:10.1016/S0378-7206(02)00014-9

Henderson-Sellers, B., Ralyté, J., Ågerfalk, P., & Rossi, M. (2014). *Situational Method Engineering*. Springer-Verlag Berlin Heidelberg., doi:10.1007/978-3-642-41467-1

Kotonya, G., & Sommerville, I. (1998). *Requirements Engineering: Process and Techniques*. John Wiley & Sons.

Lloyd, W., Rosson, M., & Arthur, J. (2002). Effectiveness of elicitation techniques in distributed requirements engineering. In *10th IEEE Joint International Conference on Requirements Engineering* (pp. 311–318). 10.1109/ICRE.2002.1048544

Ralyté, J., Deneckère, R., & Rolland, C. (2003, January). Towards a generic model for situational method engineering. In J. Eder & M. Missikoff (Eds.), *Advanced Information Systems Engineering* (pp. 95–110). Springer-Verlag Berlin Heidelberg., doi:10.1007/3-540-45017-3_9

Schnieders, A., & Puhlmann, F. (2007). Variability modeling and product derivation in e-business process families. In *Technologies for business information systems* (pp. 63–74). Springer Netherlands., doi:10.1007/1-4020-5634-6_6

Virtanen, P., Pekkola, S., & Paivarinta, T. (2013, January). Why SPI Initiative Failed: Contextual Factors and Changing Software Development Environment. In *46th Hawaii International Conference on System Sciences (HICSS)*, (pp. 4606–4615). IEEE. 10.1109/HICSS.2013.609

Vlaanderen, K., Valverde, F., & Pastor, O. (2008). Improvement of a web engineering method applying situational method engineering. In *10th International Conference on Enterprise Information Systems*, (pp.147–154). Barcelona, Spain.

KEY TERMS AND DEFINITIONS

Natural Language Model: A model built using a minimum set of structural components and completed with slightly restricted natural language text.

Process Block: A well-defined unit of process that can be reused as part of any process.

Process Commonality: The common elements of a process that facilitate the definition of a family of processes through reuse.

Process Variability: The variant elements of a process, identified at variation points, which produce deviation from the standard process.

Requirements Engineering Process: A process to construct software requirements by means of methods, techniques and tools during elicitation, modeling, analyzing and evolution of requirements.

Scenario: A representation of an observed or envisioned situation in the application context.

Situational Factor: A characteristic of the project or the application context that may be taken into account when defining an RE process.

Situational Method Engineering: A discipline that promotes the construction of methods for developing systems according to a pre-defined set of situational factors.

Variation Point: Specific location at a process where situational factors are evaluated and changes to the standard process are applied.

Digital Tools Aimed to Represent Urban Survey

9

Cristina Boido*Politecnico di Torino, Italy***Pia Davico***Politecnico di Torino, Italy***Roberta Spallone***Politecnico di Torino, Italy*

INTRODUCTION

Since the affirmation, in the Sixties, of Urban Survey as a subject in the field of drawing's disciplines, it aimed to represent a set of three-dimensional data on a plane.

This aim emerged in the European researches carried out by Pierre Lavedan and Augusto Cavallari Murat that triggered a discussion about the values characterizing the historic city, in particular the Baroque one, and the ways in which these values could be represented. Cavallari Murat and Lavedan shared that the values of the Baroque city could be represented through a concise drawing, as had happened in some famous maps of the past, in particular, that of Rome, drawn by Giovanni Battista Nolli.

In 1968, Cavallari Murat and his group of scholars published the books *Forma urbana e architettura nella Torino barocca* (Politecnico di Torino, Istituto di Architettura Tecnica, 1968), in three volumes.

The urban survey maps of Baroque Turin, drafted in scale 1: 1000, made up the third volume. They are based on the tables of the geometrical particle cadastre, drawn in 1823 by Andrea and Alberto Gatti, and on the municipal map of the time. The urban survey maps represented a selection of data collected through archival research on iconographic documents and texts, such as censuses, and aimed to become a knowledge basis for preservation and intervention on Cultural Heritage.

The morphological, typological, and stylistic data, were collected in a drawing of the ground floor, where the public spaces were distinguished to the private ones. The goal was to provide, through a map, a series of information about the three-dimensional feature of the historic urban tissue, using the so called stereometric symbols, that represented: the upper cornices, the number of storeys of the different building units, the virtual projection of vaulted and beamed coverings, the openings' axes of the main floor, the stairs as elements of vertical distribution, and the entrances to the housing units in the standard floor.

In 1974, Cavallari Murat proposed symbols for the survey of historic urban tissues, stated in the 1968's research, within the standard UNI 7310-74 - Urban Cartography. Graphic Standards, for representing historical urban tissue, mainly characterized by multi-storey buildings.

This methodology of urban survey representation can be broadened to urban tissues, not necessarily characterized by historical values.

The authors of this paper worked in the research group directed by Dino Coppo, a pupil of Cavallari Murat and are today involved in researches about the potentialities of digital tool for representing urban survey.

DOI: 10.4018/978-1-7998-3479-3.ch082

Figure 1. Philological conjectural survey of Baroque Turin. Source: Politecnico di Torino, 1968.



Indeed, digital revolution allows to represent urban tissues by 3D models that can collect a lot of information related to the buildings, becoming real data base.

Today, these models fulfil the need for update representations of urban settings, aimed both to critical studies on historical city and to manage the ongoing transformations. Moreover, these representations could become the urban scenarios for simulations and checks of master plans and architectural designs in their relationships with the built environment.

Several new tools of urban procedural modelling, BIM modelling and web resources allow generating urban 3D models.

The authors of this paper will compare the knowledge and informative capabilities of different new technologies for urban modelling, through an overview of international researches, case studies, and also some experiences personally carried out.

BACKGROUND

The representation of the environment, the city, a building, and the different aspects that define them, in an articulate relationship between tangible and intangible elements, has always been a complex matter, due to the difficulty in “narrating” and using images to summarise the various closely related aspects that make them what they are. Digital modelling combined with the latest information technology has made a fundamental contribution to creating new and constantly evolving communication scenarios which offering such a wide range of potential means and ways of representing the city as to require very careful design of the process used for the selection and communication of information.

If today’s new multimedia tools allow an immediate and interactive communication, often addressing a wide range of users, when studying a given urban or environmental reality, it is necessary to identify

forms of representation that are capable of easily and immediately transmitting urban situations and processes that can be complex, generally characterised by articulate relationships between multiple aspects, of which form or appearance are just two. The study of the city, of its being a combination of tangible and intangible elements and situations linked to its past as well as its present and even its future, configurable through images or movements generated by fruition-functional dynamics, is a study linked not only to established Urban Survey practices (Coppo & Boido, 2010) but one that is increasingly interdisciplinary, and is the basis for each representation or digital modelling, as a tool for knowledge and divulgation.

With the spread and enhancement of information technology, in recent years there have been profound changes in the access to and management of information, in spatial analysis and in graphic rendering methods. Since the 1980s, the development of GIS systems has led to formidable innovations, offering the possibility to create increasingly advanced models, using mainly data from satellite sources. The improvement in information technologies has favoured the expansion of the methods of collection, management and processing of geographical data, configuring the possibility of integrating different types of graphic rendering.

Moreover, thanks to fast and simple 3D navigation via browser and to the simplified interface of webGIS viewers, urban scale three-dimensional representations, deriving from the countless quantities of information available today, in terms of quantity, quality and variety, are considerably enhanced and widely available. The three-dimensional images of space, particularly urban space, perfected by virtual reality technologies that allow an increasingly realistic view of the environment, allow the exponential strengthening of the communicative capacity of these representations. No longer intended only for technicians working in the field but also easily understood by the general public, they allow infinite, or almost infinite, possibilities of reading, adapting to each specific need, to the point where they are almost customized. An example of this is Google Building Maker, a web application that was created in 2009, in which digital modelling was aimed at engaging the general public, allowing the individual user to play an active role in the construction of the three-dimensional image of a building to be included in Google Earth. The tool allowed the incorporation of a 3D vector representation, built with Google SketchUp modelling software, with two-dimensional raster photographs, generating scenarios suitable for depicting environmental and building features, as well as the inclusion of urban interventions. Photographs of the facades provided by the program were applied by the users on geometric 3D buildings modelled by them, and then the models were validated for inclusion in Google Earth.

This kind of representation merged 3D geometric model and photos that are products ontologically different in many respects.

Indeed, the model is a three-dimensional vectorial entity, whose faces are mostly plans, while the photos, applied on it, are two-dimensional raster entities that represent different planes projected from a point at a finite distance analogous to that perspective.

However, despite this operation was unacceptable from the science of representation point of view, it had the advantage of generating scenarios freely available for the insertion of planning and designing proposals.

The photos applied on models allowed the users to appreciate a good degree of realism of architectural morphologies and decorative styles of the context, in relation to the architectural shapes of new projects and masterplans. Furthermore, the possibility for the user to move around in a virtual reality in real-time 3D, generated a kind of dynamic photo-montage.

This method for generating urban and environmental contexts quickly became obsolete, also due to some inconsistency between the models realized by different users, and was withdrawn by Google in 2012 (Lo Turco & Spallone, 2013).

From 2017, Google Earth offers the biggest repository of photorealistic digital version of the planet. The imagery was collected by satellites (for global 2D images) and planes (for more detailed 3D data near the ground). The photos are taken from five cameras having different angulations and need to be overlapped. After preparing the photos and calibrating the colors, a depth map is created. The different maps are stitched together and a 3D mesh is realized. Finally, 3D mesh is textured with the photographs taken, producing a really photorealistic imagery.

The above described different kinds of models, used for urban analysis and planning, but also in the field of architectural design, are used as a basic support for the design of the city and its context.

Now necessary not only to represent individual situations (energy analysis, traffic, pollution, etc.), but, using specific parametric software, they become essential tools for analysis, design and visualization for city planning.

In 3D maps and models there are always more and more apps and other tools that support information and communication technologies, also in real time, with quality and accuracy. Among many, it is an example Simplex 3D models.

These applications are useful for civic innovation and development, from conception to action by combining analysis tools.

MAIN FOCUS OF THE ARTICLE

Overview

While digital modelling was used in the past to create 2D and 3D images which represented a shape or synthesized non-interactive studies on the city, today, with its use on several fronts, it has to be diachronic and questionable. It must be able to represent both present and past (just think of the reconstructions of ancient cities that have disappeared or of which only faint traces remain), and even the future (conceived, designed) to which it is possible to convey shape and attribute an image, with the possibility of adding new ideas for urban transformations (on every scale, from territorial to architectural or in detail) within the context in which they are to be proposed. This whole process requires Urban Survey methodologies and practices to work on different fronts, simultaneously considering measurable data and the elements of perceived reality (visible and not only), in an activity which, especially in its possible applications also in the design field, requires necessarily the involvement of multiple disciplinary contributions, ranging from those in the morphological and socio-economic spheres to government policies. In this scenario, which is extremely wide due to the variety of possible uses and communicative instruments that can be used, digital representation and its dissemination must be structured and designed in relation to the purposes of their use, adopting different and adequately calibrated paths of study and expression in each individual case, taking on a primary role in all kinds of communication, which can also be interactive.

Digital modelling must therefore aim to prefigure urban scenarios using different methodologies, capable of externalising the various aspects that define and characterise each area, architecture or part thereof, basing the representation on geometric-topographic matrices, which are an essential reference for what currently exists and its possible future transformations or, on the contrary, on urban situations from the past (e.g.: the incredible models that now allow us to visualize the shape of lost cities and buildings) (Brusaporci, 2010).

In this broad study process, the starting point is unquestionably defined by the data collection and processing phase, by the selection and interpretation of said data and by the choice of how to represent

them in a digital model capable of visualising the different and increasingly interactive aspects. This is how, after this phase, the various disciplinary approaches involved in the activity converge in relating through a single model, realized increasingly through the interconnection between different tools and methods (such as metric acquisition tools, georeferenced databases and a myriad of other more or less well-known 3D modelling programmes available today, some of which available also in open source format).

For the reconstruction of complex scenarios, exploiting new technologies for dynamic and interactive 3D modelling has proved and is increasingly proving to be a very useful tool, especially during the analysis and programming of interventions on an architectural and/or urban scale. Examples of this are HBIM (Heritage Building Information Modelling) in the field of restoration and requalification of the existing historical heritage, and BIM, linked in particular to new design, both of which are fundamental aids for the processes of planning, transformation and development of cities or parts of them, where it is useful to verify in advance the various aspects, not only material. The integration of the various types of information in a single model simplifies the complexity of comparison, dialogue and conversation, also reducing timing, from those concerning the survey and design phases to those of approval, through to actual operation, streamlining and accelerating each process of urban transformation and its related costs.

These new communicative approaches, which are increasingly used to gain a critical knowledge of what already exists and for the upfront testing of the consequences of transformations (large-scale or localised), in cases already experienced have led (and are leading more and more) towards the integration of information between those involved in the intervention and future users, with benefits in several directions. An interactive digital model must, in fact, improve the visibility and understanding of each phase of the process, not only for professionals but also for those who feel involved, who want to play an active role, thus promoting the implementation of a “shared interactive project”.

The visualisation of the de facto state and/or transformations of any urban reality, whatever its size, allows us to clearly understand its multiple aspects, encouraging an immediate comparison between what already exists and design, and to follow the planned process of transformation from the conceptual phase to its possible tangible and intangible effects. This strength of this new methodological approach lies in the communicative simplicity of an interactive model, because the model itself becomes the point of encounter and comparison between the different subjects who interact in the survey and transformation process (Disegnare Con, 2018). This process is characterised by the transparency of information, because the “visibility” of the intervention phases creates trust in the community, building a dialogue between the principal, the professionals engaged in the work and the general public or interlocutors involved in discussions about the future framework outlined in the model.

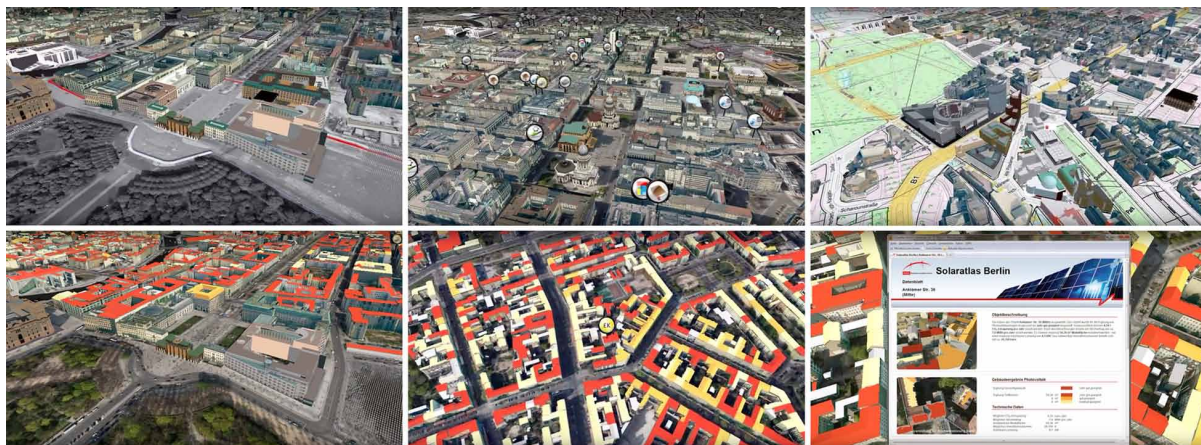
International Case Studies

The most recent experiences have shown that by increasing clarity, comparison and control of possible impacts and effects in prior simulations, applicable to every kind of operation within the city, with specific digital models, it is possible to obtain benefits. These benefits derive mainly from the correction of the operations planned, before or during the implementation phase, using the model to test the entire planning and programming process, or rather the physical, functional, economic and strategic opportunities that could derive from it, up to the point of considering those of a change in the image or liveability of a place, more closely related to people. This type of programmatic and interactive working approach is being used increasingly, and not only for large cities, due to the benefits it brings.

The possibility to interact and interface by using interactive digital modelling has meant that there are no longer separate models, created for each field of interest or investigation, but a single working tool in which to visualise and describe different operations and skills, so much so that it can also be used by non-experts in the field.

A virtuous experience is undoubtedly that begun in 2009 by the municipal administration of Berlin (Senatsverwaltung für Stadtentwicklung und Senatsverwaltung für Wirtschaft, Energie und Betriebe). Using non-commercial programs, a three-dimensional model of the city was created, based on aerial photographs of 540,000 buildings located within an area of 890 square kilometres and laser measurement of their covers. In addition, photorealistic 3D urban restoration was implemented with architectural scale models of 80 buildings, considered to be emergencies in the urban building fabric, five of which can also be visited inside (Olympic Stadium, Sony Center, Reichstag, DZ-Bank, Ostkreuz and the Central Train Station). The goal was to map the city to allow people to look at Berlin from any angle, how it is now, how it used to be and how it might change in the future. Made for non-commercial purposes, the model offers the chance to visualise recent historical transformations as well as future urban development projects, using two-dimensional and three-dimensional geodata. Using two platforms characterised by different contents and applications, *Economic Atlas Berlin* and *Berlin 3D - Downloadportal*, it is possible to take three-dimensional urban walks, consult detailed maps, view 3D buildings and retrieve specific information on the urban area of Berlin, highlighting infrastructures and public services (public transport, schools, kindergartens, scientific institutions, etc..). It is also possible to consult information on all public and private commercial properties available on the Berlin real estate market, both for those who intend to invest and for those who offer investment opportunities, directly experimenting with hypothetical scenarios of urban transformation visualised in 3D. Again, the 3D model proved to be extremely useful for visualising the city's solar potential right down to the individual building. Supported by Solar Atlas Berlin, an online atlas for calculating the potential energy yield, graphic rendering during the design phase enables verification of the suitability, adequacy and actual efficiency of a solar system to be installed on the roof of one of the city's many buildings. Flexible and versatile in adapting to different visualisations, this virtual animation of urban and partly architectural realities is one of the most interesting current representations of the city (Berlin Business Location Center, 3D City Model, 2018).

Figure 2. Berlin 3D City Model. Source: Movie by Berlin Business Location Center, from 2009

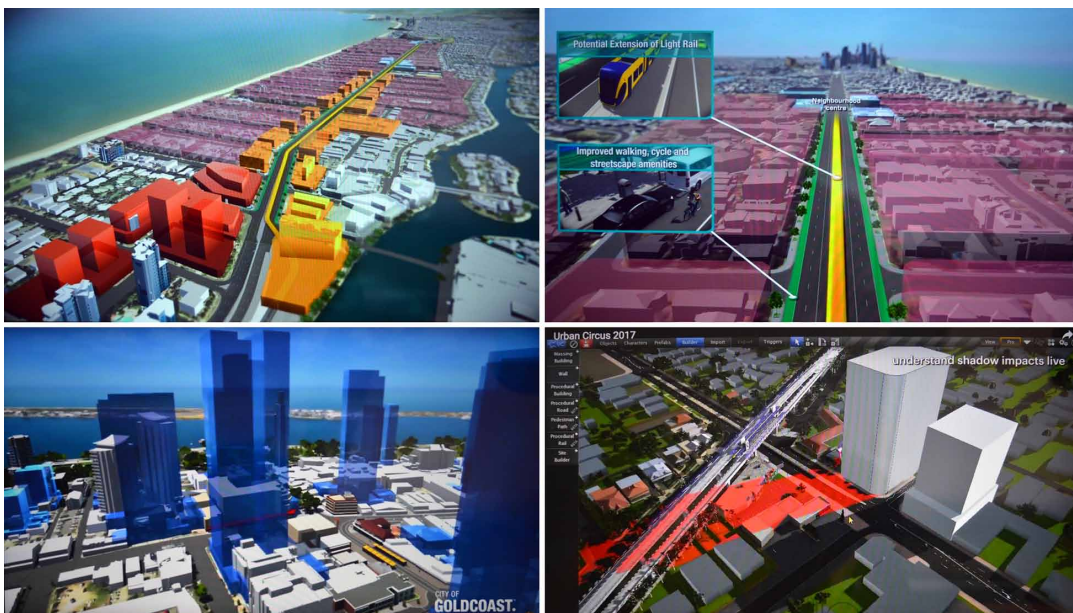


Another example related to urban scale modelling for energy verification purposes is the SeNECA project, developed as of 2015 (SeNECA - Smart and sustainABLE City from Above, 2018). It can be considered a prototype web portal capable of showing 3D visualisations of the building fabric, with a view to verifying and improving the energy efficiency of the existing building heritage, developed for the cities of Trento and Graz. Using aerial images, photogrammetrically processed to produce ortho-photos and precise three-dimensional models, compared with cadastral data and information on energy consumption and requirements, the tool offers the chance to identify where to operate to reduce urban energy consumption. The models, accessible via the web as a common basis for representing urban and energy data, and for reporting heterogeneous information and planning urbanistic and energy efficiency activities, can be consulted through mobile platforms thanks to the use of Open Source technologies.

Looking more generally at what has been achieved in recent years, with regard to urban areas we can identify models on some recurring themes, which have a greater impact on existing and future scenarios. For larger-scale modelling of cities and the territory, there are many examples of infrastructural transformations for mobility. Actually, digital models visualise the movements of the entire network, allowing us to understand the public transport service coverage of the territory, verify movements, timing, costs and the impact on the environmental image, through the construction of infrastructure (tracks, underground railways, stations, etc ...). The model, in cases that have already been tested, such as the visual communication programme “Sydney Trains”, has been a reference to help people understand, criticise, accept or reject the changes envisaged by the project, allowing corrective measures before the implementation phase (Sydney Trains – Visual Communications Program).

In other cases, such as City of Gold Coast in Queensland (2014), aimed at redesigning entire sectors of the city, the model was developed to visualise and verify the volumetric impact on the existing situation of design operations for the creation of new buildings, allowing us to understand how much, and in what way, their presence and their size would change the urban physiognomy, assessable in relation to different points of view (City of Gold Coast Draft City Plan, 2015).

Figure 3. Gold Coast, Queensland, 3D City Model. Source: Movie by City of Gold Coast Draft City Plan 2015 - Nobby to Surfers, 2015



These works often combine the study of volumes, the study of the exposure to sunshine that would occur by varying them, and the analysis of artificial lighting, to calculate and refine the lighting ratios between what already exists and the future configuration. This is carried out via visual simulation, allowing the simultaneous calculation of the energy impact resulting from the different configurations assumed, and the consequent identification of the best solutions, making it possible, using the model itself, to alter the volumes of new buildings that are planned in relation to the shadows that they would cast.

Digital modelling, with its dynamism, leads to the visualisation and understanding of scenarios of viability and architecture, linking what exists with what is proposed by the project, configuring any change that, with subtractions, additions or other variations, leads to a better understanding of every alteration, from large to small scale. An example is provided by cases concerning not only buildings but also the design of roads and links, allowing not only the visualisation of new solutions, but also the ability to interact, changing their directions, curvature, etc..., also during business meetings (sometimes in video conference), in which technicians intervene on the model, changing some elements “live”, to verify, for example, the improvement of flows or the organisation of connections with other parts of the city or territory.

Figure 4. Examples of interactive use of the 3D model. Source: Movie by Urban Circus 2017, 2017



An interesting case in which this way of working has recently been used, through the creation of a single interactive model, is that tested by the city of Melbourne. This example allows us to understand the usefulness of a single work support, which facilitates coordination between all those involved in the complex succession of phases concerning interventions on the city or on the individual building, closely related to each other.

Figure 5. Flinders Street Station, Melbourne 3D City Model. Source: Movie by Urban Circus Projects



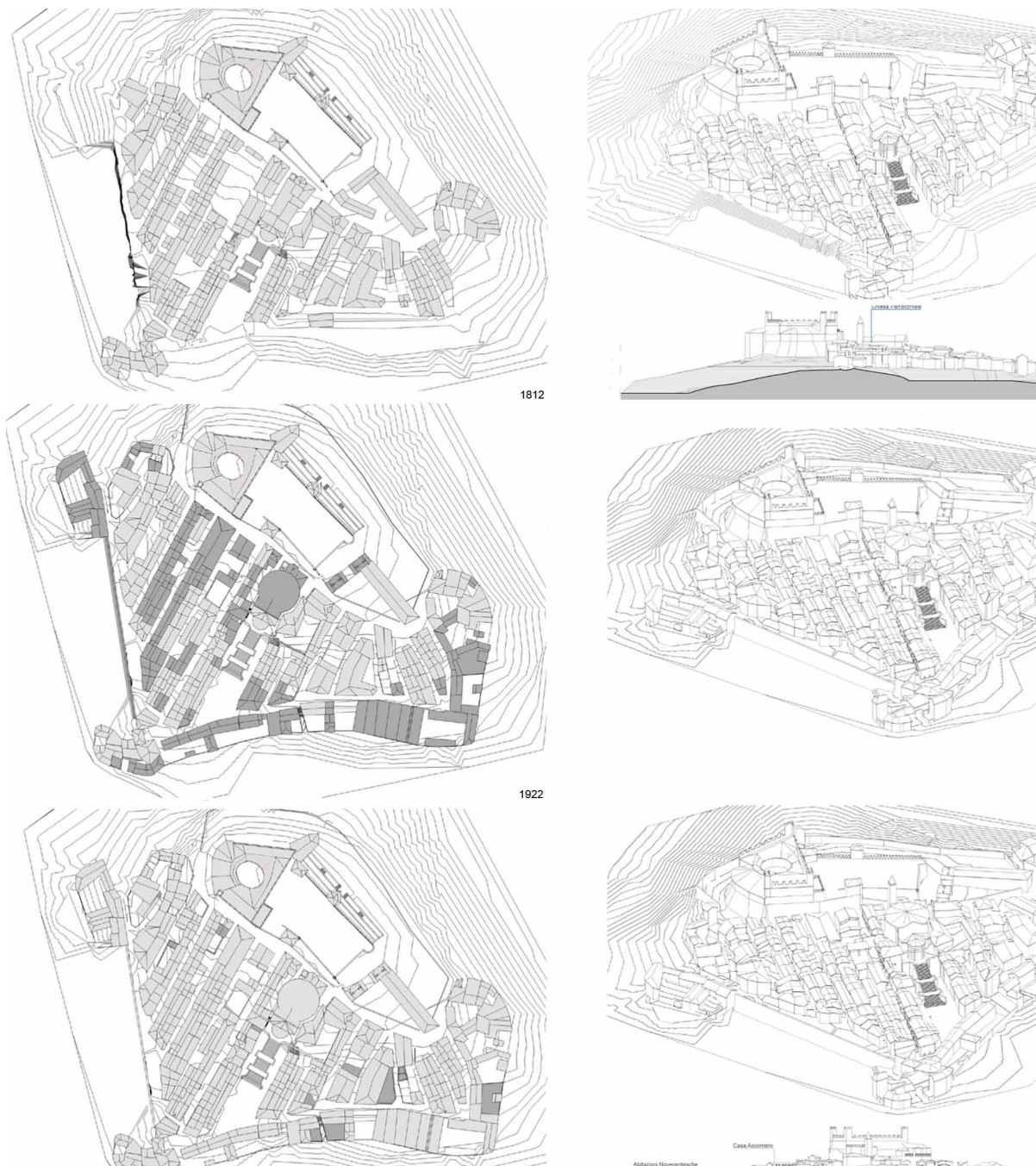
This working method is very often used in the architectural field, for single or integrated design, in which modelling through analytical steps can define and configure every aspect of the project. The model allows us, through the visualisation of the architectural aspects, to understand the configuration of the new intervention and to know the relationship with the urban context, on different levels of observation. This approach, which is also increasingly used to propose ideas for transformation to various scales of interest related to each other, is exemplified in the competition for the redevelopment of Victoria Station in Melbourne, won in 2013. The model is used to display various scenarios related to the stages of implementation, envisaging not only aspects of form and technological-constructive aspects, but also those related to timing or costs, with the possibility of continuous updating.

An Experimental Research

A case study carried out by the authors' School applied BIM technology to a diachronic and multi-scalar management of digital models, aimed to represent an urban survey.

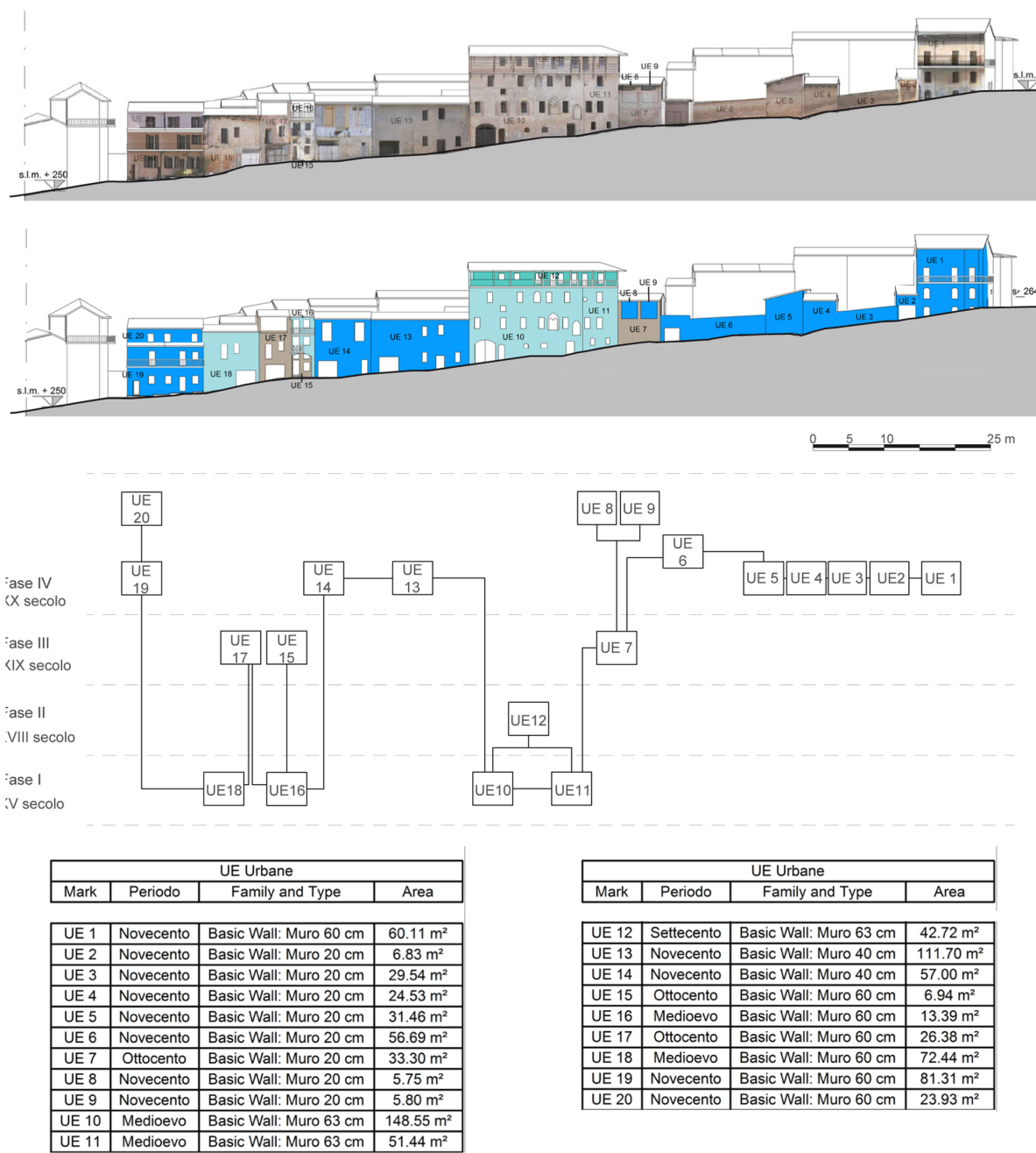
Indeed, BIM potentialities have been tested in an innovative way on the case study of a medieval settlement of particular morphology Montemagno in Asti territory (North-West Italy). It is featured by a comb-shaped layout interspersed by narrow alleys, placed on a system of three ridges. This research involved analyses at the urban, micro-urban and architectural levels, carried out by historical, bibliographic and archival research, cartographic surveys and measurements in situ. A BIM synthesis model allowed to appreciate, at different scales, the historical transformations, read with the regressive method, up to the application of stratigraphic analysis methodologies on the facade's surface of an exemplar building, the Accornero house.

Figure 6. Periodization of Montemagno (Asti) historic settlement development. In light gray the existing buildings, in dark gray the new constructions. Source: (Spallone, Piano & Piano, 2016)



The research used tools whose applications to historical buildings is currently object of study by scholars in the field of HBIM. Compared to existing studies, however, this approach used some functionality for the control of the time dimension not yet properly explored, interlacing with the archaeology of the elevations and the stratigraphic Harris Diagram (Matrix), in order to facilitate the analysis.

Figure 7. Orthophotos and Harris Diagram applied to Alley V in Montemagno. Source: (Spallone, Piano & Piano, 2016)



The case study was developed by Andrea Piano and Simona Piano as part of their Master's thesis in Architecture Construction and City, at the Politecnico di Torino, under the guide, as supervisors, of Roberta Spallone and Carlo Tosco. The interdisciplinary approach, now more than ever necessary in the investigation on Cultural Heritages characterized by a high level of complexity, such as urban and

architectural ones, thus involved the disciplines of Urban Survey, Representation and History of Medieval Architecture (Spallone, Piano & Piano, 2016).

The range of applications available today is usable and expendable on several fronts, and lends itself to expanding its fields of application thanks to new technologies that will allow us to interface more and more, even at a distance, allowing all the “insiders” involved to act by updating and changing the model, for easier coordination of work.

SOLUTIONS AND RECOMMENDATIONS

New communication languages concerning the environment and cities have led and continue to lead the field of Urban Survey to rethink its methodological approaches and, in particular, the relationship between what can be measured, what is observed and perceived, and representation (De Carlo, 2016). The application of new technologies has, in fact, triggered a cultural transformation also in the field of urban survey research, opening it up to new interconnections with other sectors and fields of application, which require new methods of investigation and communication.

The potential offered by new computer systems and programmes and by means of communication allow the involvement of an ever-growing number of people through the web, with the need to use a language that everyone, even non “insiders”, can easily understand. Urban Survey studies must therefore be proposed through new ways of expression, subverting the tried and tested specialist languages towards a more popular and less sectorial representation.

In this increasingly evolved and fascinating scenario, however, there are some controversies. In this large-scale evolutionary process, which involves many sectors, interactive digital modelling is essential in order to support complex processes of knowledge and communication, because it facilitates them by working and expressing itself through forms, images and movements, capable of simulating reality and its possible transformations, but it is important to remember that urban reality consists not only of these elements. A city, a part of it or even just a building, or its environment, are also made up of intangible elements, linked to people and their way of living and relating to one another, which can be perceived only first-hand. What and where is the boundary between reality and the model of such a complex reality?

In other words: this link between the representation of reality and future scenarios is an added value in the field of design, but how much can one “control” a communication that can convey information? How can we control use that is always adequate when the possibility to act is extended across the web? The possibility to create scenarios between the real and the imaginary highlights a fragility which we have to be careful about: representation, design (to put it simply), have always had the same potential, to create images of what exists as well as those of pure fantasy. Representation, which has always been able to create from nothing configurations that neither necessarily exist nor are true, being able to skillfully influence the “modification” of reality to convey its perception, now takes on an even more delicate role, because of the difficulty in controlling information amplified by the web, which is not always used by the right people.

FUTURE RESEARCH DIRECTIONS

The contemporary trends of the research are increasingly oriented towards the ability to automate the processes of acquisition, recognition and classification of images, as well as those of modelling territories and cities.

The acquisition systems described above, which combine satellite shots with close-up aerial shots, contribute to the automation of urban system modelling processes. Also the consolidated tools of procedural modelling and shape grammar (Parish & Müller, 2001, Müller et al., 2006) and the more recent of inverse procedural modelling, devoted to the construction of models of existing cities (Weissenberg, 2014; Cura, Perret & Paparoditis, 2018) go in the same direction.

CONCLUSION

Given the reduction in times of acquisition and modelling and an ever increasing photorealistic and detailed rendering of the reality that goes hand in hand with the multiplication of visible data, the problem of the selection and interpretation of the data that have not been established, chosen and classified a priori arises, depending on the purpose of the survey.

Moreover, as we have seen, in most cases, the representation puts together tangible and intangible elements: the latter must be visualized using a homogeneous and recognizable language. From this point of view digital tools can offer layers' overlapping, information's interweaving, multi-scalar contents.

Therefore, the problem of data selection is shifting from the moment of analysis to that of synthesis and this is one of the problematic issues that will have to engage scholars who deal with Urban Survey in the coming years.

ACKNOWLEDGMENT

Although the chapter is the result of a shared research among the three authors, it must refer to the specialist work of: Cristina Boido: the Main focus of the article: International case studies and the Solutions and recommendations; Pia Davico: the Background and the Main focus of the article: Overview; Roberta Spallone: the Introduction, the Main focus of the article: An experimental research, and the Future research directions and conclusion.

REFERENCES

- Berlin Business Location Center. (2018). *3D City Model*. Retrieved November 8, 2018, from www.businesslocationcenter.de/en/WA/B/seite0.jsp
- Brusaporci, S. (2010). *Sistemi Infortivi integrati per la tutela, la conservazione e la valorizzazione del patrimonio architettonico e urbano*. Gangemi Editore.
- City of Gold Coast Draft City Plan. 2015. Nobby to Surfers. (2015). Retrieved October 10, 2018, from <https://www.youtube.com/watch?v=SfpOIBDYcwQ>

Coppo, D., & Boido, C. (Eds.). (2010). *Rilievo Urbano. Conoscenza e rappresentazione della città consolidata*. Alinea.

Cura, R., Perret, J., & Paparoditis, N. (2018). *A state of the art of urban reconstruction: street, street network, vegetation, urban feature*. Cornell University. Retrieved June 30, 2018, from <https://arxiv.org/abs/1803.04332>

De Carlo, L. (Ed.). (2016). *Metamorfosi dell'immagine urbana. Rappresentazione, documentazione, interpretazione, comunicazione*. Gangemi Editore.

Disegnare Con. (2018). Co-Design: digital tools for knowledge-building and decision-making. *Disegnare Con*, 11(20).

Lo Turco, M., & Spallone, R. (2013). Urban transformation scenarios for the representation and dynamic control of new design interventions. *Planum. Journal of Urbanism*, 27(2), 1–9.

Modellazione 3D a scala urbana per l'efficientamento energetico di edifici. (2016). *Rivista GeoMedia*. Retrieved October 15, 2018, from https://rivistageomedia.it/201608039762/bim-cad-gis/seneca-smart-and-sustainable-city-from-above-modellazione-3d-a-scala-urbana-per-l-efficientamento-energetico-di-edifici?acm=_421&idU=2

Müller, P., Vereenooghe, T., Wonka, P., Paap, I., & Van Gool, L. (2006). Procedural 3D reconstruction of Puuc buildings in Xkipche. In *Eurographics Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*. Geneva, Switzerland: Eurographics Association.

Parish, Y. I. H., & Müller, P. (2001). Procedural modeling of cities. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. Los Angeles, NV: ACM.

Politecnico di Torino. Istituto di architettura tecnica, research coordinator Augusto Cavallari Murat. (1968). *Forma urbana ed architettura nella Torino barocca: dalle premesse classiche alle conclusioni neoclassiche*. Torino, Italy: UTET.

SeNECA - Smart and sustaiNable City from Above. (2018). Retrieved November 8, 2018, from <http://seneca.fbk.eu/home>

Spallone, R., Piano, A., & Piano, S. (2016). B.I.M. and cultural heritage: Multi-scalar and multi-dimensional analysis and representation of an historical settlement. The case study of Montemagno, a New Village in Piedmont. *DisegnareCON*, 16, 1–12.

Sydney Trains – Visual Communications Program. (n.d.). Retrieved October 8, 2018, from <http://urban-circus.com.au/project/sydney-trains/>

Weissenberg, J. (2014). *Inverse Procedural Modelling and Applications* (Unpublished doctoral dissertation). ETH Zurich.

ADDITIONAL READING

Lipp, M., Scherzer, D., Wonka, P., & Wimmer, M. (2011). Interactive Modeling of City Layouts using Layers of Procedural Content. *Eurographics*, 30(2), 1–10. doi:10.1111/j.1467-8659.2011.01865.x

Müller, P., Wonka, P., Haegler, S., Ulmer, A., & Van Gool, L. (2006). Procedural modeling of buildings. In *ACM Transactions on Graphics* (3 ed., Vol. 25, pp. 614-623). Retrieved July 30, 2018, from doi:10.1145/1179352.1141931

Nishida, G., Garcia-Dorado, I., Aliaga, D. G., Benes, B., & Bousseau, A. (2016). Interactive Sketching of Urban Procedural Models. In *ACM Trans. Graph*, 35 (4). Los Angeles, USA: ACM. doi:10.1145/2897824.2925951

Toschi, I., Nocerino, E., Remondino, F., Revolti, A., Soria, G., & Piffer, S. (2017). Geospatial data processing for 3D city model generation, management and visualization. In *Proceedings of the The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-1-W1, 527-534. Hannover, Germany.

Valencia, J., Muñoz-Nieto, A., & Rodriguez-Gonzalvez, P. (2015). Virtual modeling for cities of the future. state-of-the art and future challenges. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5(W4), 179–185. doi:10.5194/isprsarchives-XL-5-W4-179-2015

KEY TERMS AND DEFINITIONS

3D Geometric Modelling: It is a modelling method based on points, lines, surfaces and volumes.

3D Modelling: 3D modelling is the process of developing a mathematical representation of any three-dimensional surface of object via specialized software.

3D Parametric Modelling: It is a modelling method that allows to maintain consistent relationships between elements as the model is manipulated.

BIM Modelling: Building information modelling (BIM) is a digital representation of physical and functional characteristics of a facility. A BIM is a shared knowledge resource for information about a facility forming a reliable basis for decisions during its lifecycle; defined as existing from earliest conception to demolition. BIM involves representing a design as combinations of objects that carry their geometry, relations, and attributes.

Database: A complex combination of structured data in tables that relate to one another, allowing the interactive management of information with automatic implementations, quick updates and specific interrogations.

GIS: Geo-referenced information system which allows the relation and organisation of geo-referenced data, favouring the update, enhancement, analysis, manipulation, and re-elaboration in map form.

Open Source: The term indicates software that is not protected by copyright, the source code of which (this being the part of the software that computer programmers can manipulate to alter the operation of a program or an app) is issued with a licence which makes it open to alteration or improvement by users.

Shape Grammar: Invented by Stiny and Gips (1972), it is an algorithmic system for creating and understanding designs directly through computations with shapes.

Addressing the New Pragmatic Methods in Urban Design Discipline

Hisham Abusaada

 <https://orcid.org/0000-0001-6530-7714>

Housing and Building National Research Center (HBRC), Egypt

Abeer Elshater

 <https://orcid.org/0000-0002-5061-6861>

Ain Shams University, Egypt

INTRODUCTION

Almost three decades before the beginning of the third millennium, the impacts of the “digital age” appeared and dominated most areas such as theories of journalism, teaching, foreign language education, new epistemologies, digital media, and digital fluency. Consequently, before the beginning of the twenty-first century, there were impacts on things like the faces of buildings, the design of public urban spaces, urban industry, urban mobility, urban policy, and integrating telecommunications. These effects have been apparent in most studies in the areas of planning and urban design and have become influential in the accuracy or weakness of the results, mainly in developing the cyber-cities, smart cities, smart growth, and digital urbanism. These influences have reached the point where most of the scientific forums have begun insisting on the application of new, innovative methods that rely on digital technologies along with developed conventional methods.

The primary goal will be to understand the importance of using research methods appropriate to the subject matter of the research and the desired results. Also, it intends to drive the researchers to develop their vision by using the traditional methods and seeking to integrate them with advanced technology research methods in their related work. This goal includes two objectives. The first is to highlight the importance of using both new and conventional methods together. The second is to present these selected analytical methods in a way that allows the readers to search for the other purposes associated with their research.

This chapter addresses new methods that display research problems which are strongly linked with two issues: urban form (UF) and spatial analysis (SA). It focuses on how to use these methods in an urban design discipline (UD), based on the dominant science and new analytical approaches, such as “virtual world design,” “spatial planning,” “geoinformatics,” “urban sensing,” and “a three-dimensional model of the city.” It reveals the methods that help the designer to measure people’s sensation in cities. These methods try to balance between two fundamental issues. The first is collecting data about city places in each survey about people’s behavior. The second is exploring the reliability and validity of these methods and measures by pragmatically applying them to the analysis of real-world problems.

The research methodology focuses on a bibliographic study, including application examples, based on a considerable number of academic research papers and official reports, publications dealing with some conventional and new pragmatic methods in urban planning and design discipline, as well as in the area of environment design. This study will first display the background theoretical approaches related

DOI: 10.4018/978-1-7998-3479-3.ch083

to the areas of planning and design, as an introduction to clarify the research methods that underpin the central theme.

Notably, this study is a bibliographic review of some new pragmatically methods in urban planning and design discipline. The organization of this chapter starts with a background for the key issues, and one episode has a connected view that includes six methods in two stages. The first displays two methods for measuring the visual aesthetic dimension, and the second consists of four methods for measuring the affective response approach. The conclusion states the definitive view and includes a proposal for future research.

BACKGROUND AND SETTINGS OF URBAN FORM AND SPATIAL ANALYSIS

Urban form is one of the most critical concerns of urban design. It focuses on the relationship between people and place in the built environment, with particular attention to the relationship between built-up mass and urban spaces. For decades, the specialists have devised several dimensions which govern these bilateral relations people-place and built-up mass-spaces. They are functional, social, economic, environmental, behavioral, aesthetic-visual, cognitive-perceptual, and temporal dimensions. Spatial analysis of urban spaces is one of the conventional analytical methods that urban design has used for measuring and analyzing the urban form, which was extended, based on Kevin Lynch's (1960) famous theory "the Image of the City".

In 1960, Kevin Lynch focused on the concept of the 'cognitive map' and conducted his first real-life experiments in Boston and Los Angeles, when he asked the residents to sketch out and bring back their daily routes from their memories. He presented the "way-finding" strategy, legibility and imageability to be a theoretical base for various studies. Afterwards, this strategy is transformed to be a tool, an approach, a quantitative analytical method, and a methodology. In these, they examine the spatial system and the configuration of the layout of the city. Furthermore, it explores the effects of this configuration on the organization of the uses within the city and the relations with the entire system.

Jerry Weisman (1981) and GD Weisman (1987) used the term of legibility based on the wayfinding method, to identify the ability of users to find their ways in the layout of the student' building and in housing for elderly people. John Peponis, Craig Zimring, and Yoon Kyung Choi (1990) developed this method to achieve a *more rigorous analytic description of building layout and exploration paths that exhibit their own pattern* (Peponis, Zimring, & Kuang Choi, 1990, p. 550). The wayfinding method has been used to study the structural properties of the building layout depends on developed the techniques of syntactic analysis, at University College London. Distinctly, the most common studies starting from Kevin Lynch's (1960) *The Image of the City*, Jerry Weisman's (1981) *Evaluating Architectural Legibility Way-Finding in the Built Environment*, Romedi Passini's (1984) *Wayfinding in Architecture*, GD Weisman's (1987) *Improving Wayfinding and Architectural Legibility in Housing for the Elderly*, John Peponis, Craig Zimring, and Yoon Kuang Choi's (1990) *Finding the Building in Wayfinding*, Paul Arthur and Romedi Passini's (1992) *Wayfinding: People, Signs, and Architecture*, Susan Hunter's (2010) *Design Resources: Spatial Orientation, Environmental Perception, and Wayfinding*, Gesine Marquardt's (2011) *Wayfinding for People With Dementia: The Role of Architectural Design..*

In the vein, according to Michael Mehaffy (2008), Christopher Alexander is one of the pioneers of the design method movement that aims to move systematically through sequential steps in an organized design process. The movement's response to the complexity of technical problems in the post-World War II era is fuelled by the development of computer science and programs. Instead, the new method should

incorporate an “argumentative process” within a network of issues. In the vein, according to Michael Mehaffy (2008), Christopher Alexander is one of the pioneers of the design method movement that aims to move systematically through sequential steps in an organized design process. The movement’s response to the complexity of technical problems in the post-World War II era is fueled by the development of computer science and programs. Instead, the new method should incorporate an “argumentative process” within a network of issues. After the first generations of design methods (the problems) and the second generation (argumentative process), the third generation (refutable hypotheses) emerged in 1980 with respecting the “plurality of views”. The third generation leads to the Christopher Alexander’s theory of pattern languages (1977).

The methods of quantitative analysis of urban spaces were introduced using new theories, including Bill Heller’s theory of the space syntax (SS) (Hillier, 1996). It describes layouts regarding the pattern of connections between spaces based on the “kinetic perception.” It was based on the idea that gathering information about the city comes through the movement and mobility within its streets (Jabbari, Fonseca, & Ramos, 2017).

The SS has emerged as a theory for the social use of space and virtual world design from since the early 1970s. Its applications are in the field of urban design in the “social field of vision,” and the cultural aspects of human behavior and communication in social spaces. It focuses on the study of spatial interrelationships between people’s interactions to explore the “social distances”, and distinct levels of intimacy (Hillier, 1996; Hillier & Hanson, 1997; Morello & Ratti, 2009). The SS (Batty, 1991, 1997; Hillier & Hanson, 1984, 1996; Peponis & Wineman, 2010) started as a theory for studying the configurational properties of cities.

Since 2000s, the innovative approaches have emerged to address issues related to the human soul and their emotional responses, and the human sensory assessment using sensor technologies concerning humans’ perception of spaces and emotional reactions in real time. The information related to emotion was used in the urban planning of the concept of smart city (Choudhury, Pradhan, & Kar, 2016; Matei, Ball-Rokeach, & Qiu, 2001; Sagl, Resch, & Blaschke, 2015). Susan Hunter (2010) sees that “researchers studied how the layout of built structures influenced human emotions and movement” (p. 2).

In 2015, Peter Zeile, Bernd Resch, Linda Dörrzapf, Jan-Philipp Exner, Günter Sagl, Anja Summa, and Martin Sudmanns, in their research entitled *Urban Emotions—Tools of Integrating People’s Perception into Urban Planning* used conventional methods, such as “mental maps and psycho-geography” with “the field of digital emotion-based urban tagging” (Choudhury et al., 2016, p. 27). Moreover, Bernd Resch (2013) Choudhury et al. (2016) believe that the term urban motion used the concepts of “people as sensors” and “urban tagging.” Also, the specialists could understand the concept of smart city through new bio-statistical and sensor technology. Urban motion depends on the spatial planning, geographic information systems, concepts of computer linguistics, sensor technology methods, and real-world data. It uses:

1. the spatial monitoring approach based on sensor driven data gathering and human sensory assessment. Monitoring includes deductive monitoring by observation and inductive monitoring through crowd sourcing processes.
2. The Psycho-physiological monitoring approach for measuring the arousal of the citizens, based on the body reaction. It uses the gathering of bio-sensor, bio-statistical, and physiological data.

PRAGMATICAL METHODS: CONVENTIONAL ALONGSIDE TECHNICAL

This section discusses two sets which are related to the concept of spatial analysis. The first set focuses on the visual-aesthetic dimension which handles several terms and concepts, such as wayfinding, imageability, and walkability. It addresses two urban planning and design analytical methods, starting from the 1960s. The second set focuses on the fields of urban design based on earlier science and new analytical approaches, such as virtual world design, spatial planning, geoinformatics, urban sensing, and a three-dimensional model of the city. Modern technological innovations have helped to develop scientific research methods.

Some of these innovations are geographic information science, communication techniques, modern wireless telecommunications (i.e., general packet radio service), the third-generation mobile phone system, universal mobile telecommunications system, wireless local area network, actuators, radio-frequency identification tags, remote sensing, mobile phones, Internet technologies, developed mobile technologies, ubiquitous mobile computers, GPS trajectories (Resch, Summa, Sagl, & Zeile, 2015; Choudhury, Pradhan, & Kar, 2016). They discuss the affective dimensions. For instance, it stimulates human feelings when someone visits a place for the first time, which is known as the “urban affective atmosphere.” In this regard, the first investigation should be about the meaning of addressing or using “new methods” to display research problems which are strongly linked with the issue of “urban emotions” and affective atmosphere. Particularly, the issues of urban emotions have been discussed repeatedly in the fields of environmental psychology, urban sociology, and urban planning.

This section reviews six methods dealing with the emotional information in the context of urban planning and design starting from the year of 1960. These methods can be classified as follows 1) wayfinding (1960), 2) Space Syntax (1970), 3) Human Sensory Assessment (2000), 4) Urban Sensing: people as sensors (2008), 5) Crowdsourcing approach cognitive appraisal (2008), and 6) The Internet of Things (IoT) (2010).

Two Methods for Measuring the Visual-Aesthetic Dimension

Based on people’s intuitive ability, four components of controlling the aesthetic appreciation include:

1. **Sense of Rhythm and Pattern:** Large affinity, similarity, contributing to character, and identity.
2. **Appreciation of Rhythm:** Stricter repetition: contributing to visual pleasure.
3. **Recognition of Balance:** Visual scene, and order related to harmony.
4. **Sensitivity to Harmonic Relationships:** A coherent whole (Carmona, Heath, Oc, & Tiesdell, 2003, p. 131). It includes two methods: The wayfinding approach (Lynch, 1960) and the SS.

The Wayfinding Approach: The First Method

This method is followed by the real-world design through involving people in monitoring the city components. It depends on the concept of cognitive maps that Kevin Lynch devised in 1960 to perceive “the image of the city.” These maps include five key elements: Districts, landmarks, paths, edges, and nodes. Also, the maps include people’s feelings of the place, and this approach evolves to include three criteria: Legibility (clarity), imageability (visual image), and wayfinding (direction). Legibility is easy for anyone to recognize the parts of a city and can also organize them in “a coherent pattern” (Lynch, 1960, p. 2). This means that the city is organized in a way the person can easily read the cohesion of

the city parts. The legible elements make the city easier for the viewer to interact with the context and to establish a relationship that allows him/her to come to the place each time and reach his/her goals quickly. This helps him/her not to lose the direction, track, or destination to reach. Each time, through this reading, the viewer has a visual experience that makes him/her remember the place through an event or situation, or even through the building itself. For any viewer, the imageability explains the degree of clarity of the mental image of the city, it is a physical object, reflected from its character and quality. Whenever the quality and character of the city is reliable, then the opportunity to evoke a strong image in the observer has high probability. The shape, colour, and arrangement are involved in the creation of clear and robust mental images of the environment. Also, it helps “*the making of vividly identified, powerfully structured, and highly useful mental images*” (Lynch, 1960, p. 9). Kevin Lynch presented wayfinding as a tool to create the image of the city. Through reshaping cities, Lynch’s goal is to give a form that organizes the context (Lynch, 1960, p. 90) that depends on “*how well people are able to find their way to their particular destination without delay or undue anxiety*” (Peponis, Zimring, & Kuang Choi, 1990, p. 561).

Several empirical ways allow to analyse the wayfinding performance: a) Self-report, b) the ability to locate places on a plan of the premises, c) the proportion of subjects which are able to be reached of a particular destination (quantify), and d) direct observation, but it is relatively rare (Peponis et al., 1990, p. 560). Constantly, people tend to orient themselves towards urban spaces that are easily accessible. These urban spaces should be legible by the following three elements. The first is the “structural properties of buildings layout. The second is the configurational properties, rather than the merely rely on the landmarks, signs, or other cues. The last is the data that should reflect the distinction between the general understanding of layouts and the search for specific destinations (Peponis et al., 1990, p. 556). The configuration properties refer to several aspects which are:

- The way in which spaces are related to one another.
- The overall pattern that constitute of these spaces.
- The overall pattern that emerges from pairwise connections, rather than elements or single connections taken by themselves (Peponis et al., 1990, p. 558).

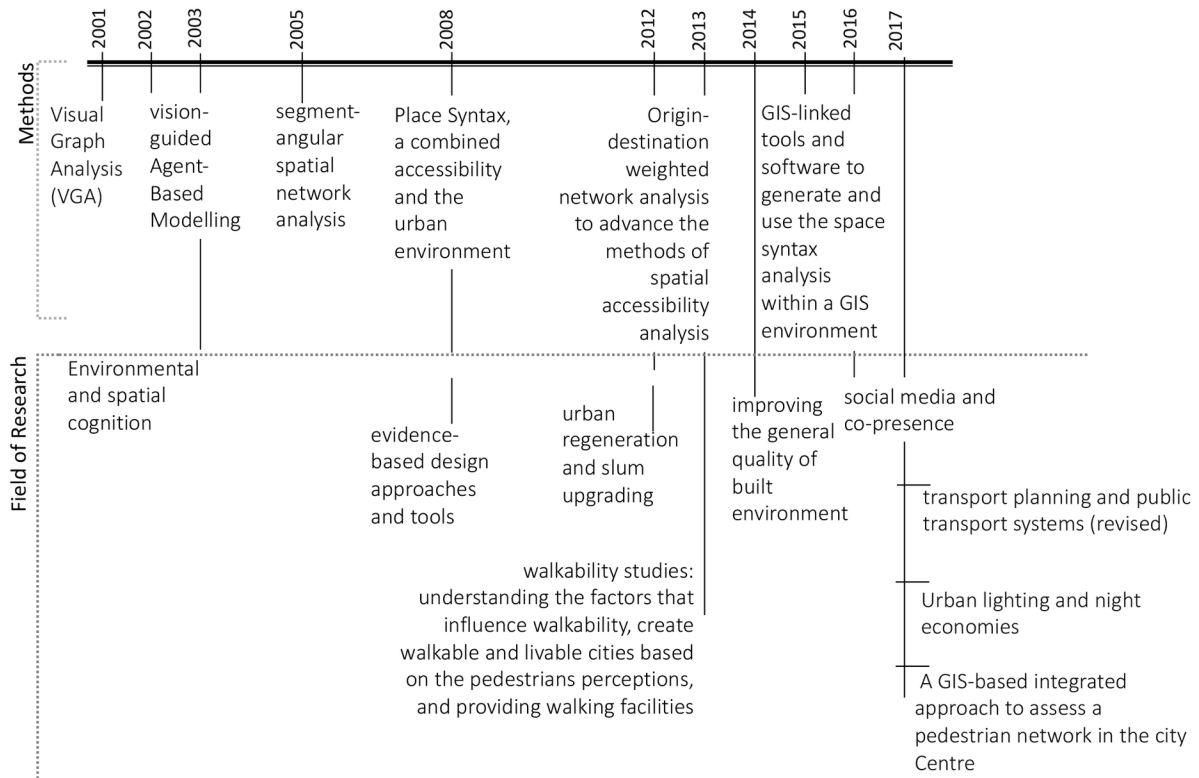
In 1992, at the University of Montreal, Paul Arthur, as an architect, and Romedi Passini, as an environmental psychologist, developed the term “wayfinding” to be used in a framework of collaboration of cognitive architecture (Arthur & Passini, 1992). The wayfinding is the environmental stimuli (i.e., intersection, stairs, and billboard) that orients people’s choices and behaviours (e.g., turn right, go up, and look for information) (Arthur & Passini, 1992, p. 31). The wayfinding becomes a tool to the information-gathering and decision-making processes. It used to orient and let people move easily through a space as Susan Hunter (2010) describes by “how people get from one place to another” (p. 1). Moreover, in this method, the cognitive space is classified as a topological space; people identify the relative locations, rather than the direction or distance (Hunter, 2010, p. 3). The result is that people’s cognitive maps about the surroundings are articulated based on the information which comes from their knowledge (Hunter, 2010, p. 3).

Space Syntax: The Second Method

This theory deals with the city as a living being. Also, it is a spatial analysis of many urban spaces that provide reliable information in the observers’ perception, whether they are residents or visitors. The

idea of this method is based on the “kinetic perception,” or what we can call the “sequential movement-cognitive.” People’s sense of the city is done by perceiving each segment as a prelude for realizing the whole, and this realization is carried out through multiple space trips within the city. The viewers move through the spaces and are aware of the place, then they move to another space. Each movement has different information and different awareness; this flow continues until people have a full understanding of one trip, and then of a series of tours. This method was based on the idea of gathering information about the city comes through the movement and mobility within its streets (abbari, Fonseca, & Ramos, 2017, p. 23). The observations that the movement arises create the mental image of the city. Therefore, this theory develops the method of axial maps, which rely on the analysis of roads network, the relations with each other, and land use overlooking them. In this theory, the focus of the urban analysis is the “city spaces and their organization” to give approximately 60–80% of the pedestrian and vehicular movement pattern (Saif Haq & Girmay Berhie, 2017, pp. 5-6). The SS method is used to analyse the spatial configuration of the city based on “how the urban system functions” (Kayvan Karimi, 2017, p. 1). These methods are used in many fields (Jabbari et al., 2017, p. 23; Kayvan Karimi, 2017, p. 1), as shown in (Figure 1).

Figure 1. Space syntax timeline of methods and field of research. Source: the authors



In the early 1970s, the University College London presented the SS theory to discuss the social use of space, followed by the fields of virtual world design. In architectural and urban design, the SS presents the applications to study spatial interrelationships between people’ interactions, particularly in the “so-

cial field of vision,” and the cultural aspects of human behaviour and communication in social spaces. Mainly, the SS allows to explore the “social distances” that enable diverse types of human activity and distinct levels of intimacy in the interrelationship between human beings (Hillier, 1996, p. 64; Hillier & Hanson, 1997; Morello & Ratti, 2009, p. 838). For Hillier and Hanson (1997), this concept becomes an approach to examine the reasons of why buildings and the built environment acquire this spatial form. This approach searches about why buildings and built environments were as they were and occupied only a small corner of the theoretically vast field of architectural and urban possibility. Notably, the initial studies of this approach dealt with “the organic settlements” and “vernacular buildings”; it tried to show the formal, spatial, and functional forces that generate their characteristics of the spatial forms. The SS as an approach focuses on the “perceptual dimension” of urban spaces, through a continuous observation journey in the place for a specific period of time and has to do with the nature of the structural structure of the target location. It is based on providing “a digital image of the city” and its area is “remote sensing” and “spatial information science.”

In 2008, the SS defined as “a methodology,” studied two aspects of the spatial system of the city: a) The connective relationships for the entire system, and b) the global connective properties of each element. Also, it was used as “a tool” to examine the nature and function of the activity that should be found in a specific place through its degree of relevance to the local context, as well as its relevance to the public context (Mehaffy, 2008, p. 61; Wineman & Peponis, 2010, p. 90). This process includes two phases: Phase I-Indicators of structural characteristics analysis and Phase II-Axial maps and lines analysis.

Phase I-Indicators of Structural Characteristics Analysis. This method reviews the characteristics of the structure to the case study starting the city levels, through a combination of indicators. It focuses on the following six indicators:

1. **Intelligibility:** It is an indicator that shows the degree at which people can see some places directly and create a subjective experience without any effort. It implies that places cannot be seen and tested for the same place in the same way.” Any system that can be extrapolated” means that its spaces are well connected and integrated. Conversely, the area that cannot be read is connected, but not integrated. Thus, the area does not drive the visitor well (Hillier, 1996, pp. 129, 171; Morello & Ratti, 2009, p. 845; Wineman & Peponis, 2010, p. 90).
2. **Connectivity Index:** It is concerned with the extent of communication between the components of the project. It is the study of the patterns of connections through the relationship of each spatial unit or line to its immediate adjacent. Also, it measures the flexibility of movement that a specific space should provide through the number of axes of movement linked to each other.
3. **Synergy Index:** It is the interaction and cooperation of two or more elements to create a joint effect, greater than the sum of their separate effects. This indicator measures the flow of traffic, based on the correlation coefficient, which shows the relationship between local and global integration. This indicator is concerned with measuring the residents’ privacy and isolating them from visitors through the urban structure.
4. **Integration Index:** It is the relationship of each spatial unit or line to the entire set of other units or lines that constitute the spatial system that is studied and measured. It measures “the relative position of space or axial line vis-a-vis the general pattern of the building configuration” (Peponis et al., 1990, p. 567). This indicator measures the extent to which the components of the study area are compatible with each other and for their inhabitants or visitors, according to changes in transitions and change over time. Technically, “integration” and “choice” indicate the complex values for

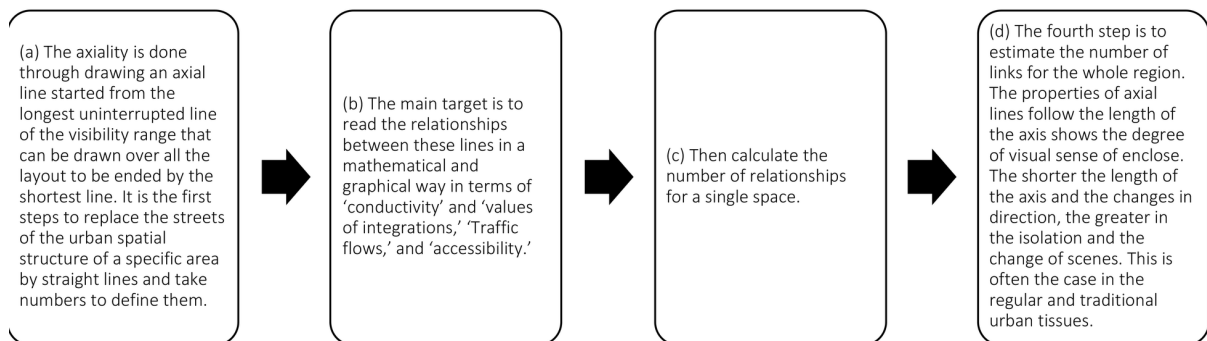
each “unit area or space.” It means a measure of “closeness” and “how close space is, on average, from all other spaces in that network” (Saif Haq & Girmay Berhie, 2017, p. 6).

5. **Spatial Openness Index:** It is used to reinterpret the two-dimensional city image through a three-dimensional analysis; it depends on the relation between social dimension and the built environment (Dalton & Bafna, 2003; Morello & Ratti, 2009). The index is defined as (a) the volume of the part of a surrounding sphere which is visible from a given point of view, and (b) the net volume of open space. It is determined by the quality of perception and comfort based on the perceived density (Morello & Ratti, 2009, p. 842) to consider the notion of “intelligibility” that explains the quality of an environment (Dalton & Bafna, 2003, pp. 1-59).
6. **Intensity Index:** It measures the intensity of the use of the axes by the resident population and visitors, and, in theory, the more axes of movement associated with each other, the less density on each axis.

Phase II-Axial and Convex Maps and Lines Analysis by Using Iso-Visi-Matrix. For the SS, based on the concept of the convex space partitioning (the simply axial mapping), the “continuous space” is converted into “a connected set of discrete units.” Hillier (1984, pp. 97-98) generated the convex map of the spatial structure that is partitioned into a set of “fewest and fattest” convex spaces. The convex map is the most elementary unit of analysis that comprises the fewest and fattest convex spaces. It has two-dimensional extensions directly visible from each segmental point. Convex spaces are the largest units that can be fully perceived at one time within a building and can be taken to represent the “local constituents” of a layout (Peponis, Zimring, & Kuang Choi, 1990, p. 566).

This method discusses the implications of two issues. The first is the prediction of pedestrian traffic in residential areas and their relation to the motorised traffic through accessibility, traffic flows, and wayfinding. The second is the patterns of residents’ and visitors’ collective behaviour to estimate how much crime can occur in the area. The steps of drawing an axial and convex map are listed in Figure 2. The significant consequence of this map is showing how vital each “vacuum” is to the structure that gives the affinity or isolation to the population’. This way may be useful when comparing two or more streets through reading the clarity, regular communication, and integration in deciding the feelings of belonging to the place, familiarity, fear, and safety.

Figure 2. Steps of space syntax. Source: the authors



Tandy (1967) presented the notion of “isovist” in the field of landscape geography. An isovist is the set of all points visible from a given vantage point in space and concerning the environment (Benedikt,

1979, p. 47). The shape and size of an isovist are subject to change with position. Numerical measures that quantify some salient size and shape features are proposed. It is known as (a) “the field of view, available from a specific point of view,” and (b) the area not in the shadow cast from a point light source (Morello & Ratti, 2009, p. 839). Also, “an isovist is the sum of the infinite number of lines of sight (or axial lines) that pass through a single point in space (usually at eye level) and found in the same plane (usually parallel to the ground level)” (Morello & Ratti, 2009, p. 839). Isovist is the volume of visible space from a given point in the area, together with a specification of the location of that point. Sets of isovists and isovist fields form an alternative description of the environments. This method is related to the behavioural and perceptual studies in the field of architecture. Furthermore, it focuses on several issues, which are the areas of view control, privacy, defensibility, dynamic complexity, and spaciousness (Benedikt, 1979, p. 47).

The spatial units for syntactic analysis are based on the visibility polygons or isovists, which is “the field of view” available from a specific vantage point; a horizontal slice through this field of view is then calculated, usually taken at eye height and parallel to the ground plane (Benedikt, 1979). The two characteristics describe each polygon as, on the one hand, a visibility polygon covers all points that can be linked to a given root point by a visibility line. This line is not interrupted by any boundary. On the other hand, the vision thresholds that share the polygon of vision about all other locations is in the same pattern (Wineman & Peponis, 2010, p. 89).

The software is used to fill all navigable spaces within the area of study with a grid of vantage points and to generate visibility (eye-level isovists) or accessibility (floor-level isovists) polygons from each of these locations. Each of these polygons that are created from the grid of vantage points can be characterised by a series of properties that include area, perimeter, minimum, mean, and maximum radial length. Once the grid polygons have been calculated, the researcher can examine the relationship between each polygon generating point and every other generating point to develop a set of “syntactic” measures (Wineman & Peponis, 2010, p. 89). When visibility polygons, or isovists, are drawn at eye level, they capture the objective properties of the visual field as structured by the affordances of the environment (Wineman & Peponis, 2010, p. 89). A three-dimensional matrix of urban visibility (i.e. a “3-D isovist”) “defines the 3-D field of view, which can be seen from a vantage point with a circular rotation of 360 degrees and from the ground to the sky” (Morello & Ratti, 2009, p. 842).

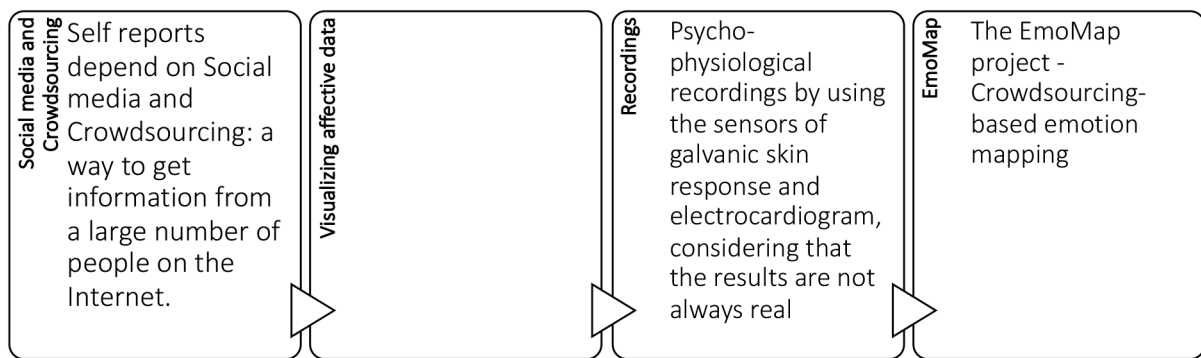
Four Methods for Measuring the Affective Response

Similar to James Russell (2003, 145–149), the current work discusses concepts of the primary emotional impact such as the core affect, perception of affective quality, and attributes. It describes many of the everyday feelings. The effect is the simplest raw (non-reflective) reaction to the feelings and mood, which is described as a neurophysiological state. In this, the affective quality is a property of the stimulus. The attributed effect covers the reactions, liking, displeasure, motives, and empathy. Affective response is presented by a person, either perceived as neutral, moderate, or extreme. When a person is responding to a particular situation, the response is twofold as emotional and conscious experience. The first is the emotional experience when it is affected by the moderate or extreme; it can be consciously experienced as pleasant or unpleasant and form the basis of an emotional experience. The second is the conscious experience. When the effect is milder, it influences the conscious experience and the behaviour in a less direct way. In this case, the effect is experienced in the background—as the property of the environment. Both types of experience can diagnose the environment, but the first is complex and difficult to measure

and is linked to mood and feelings (Klettner, Huang, Schmidt, & Gartner, 2013, p. 66), (Resch, Summa, Sagl, & Zeile, 2015, p. 202).

This approach is diagnostic of the affective responses to space and examines the emotional qualities of environments (i.e., hectic, exciting, or beautiful). It includes an innovative approach to collecting the affective responses towards space (Klettner, Huang, Schmidt, & Gartner, 2013, p. 67), (Figure 3). In this approach, the process includes two methods: a) deductive—the observation of phenomena over a time span, and b) inductive—an expression of crowdsourcing processes in the bottom-up planning mode (Zeile, et al., 2015, p. 907). ARA includes four methods: human sensory assessment (HSA), urban sensing (US), Crowdsourcing, and The Internet of Things (IoT) (Abusaada & Elshater, 2019a).

Figure 3. The affective responses to space that examines the emotional qualities of environments. Source: the authors



Human Sensory Assessment: The Third Method

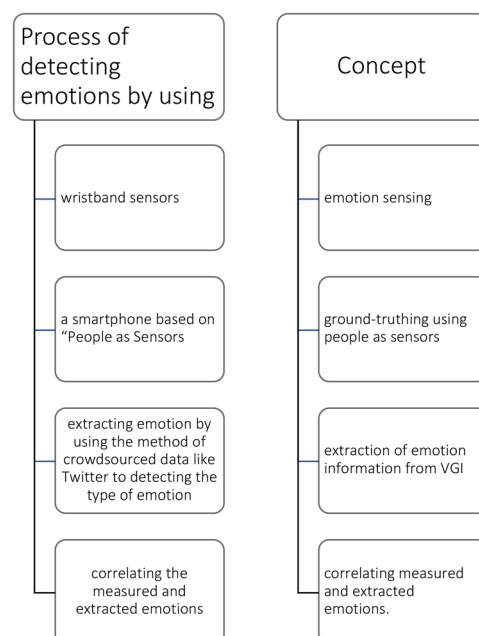
The human sensory assessment has appeared at the beginning of 2000 as a subjective method of the data collection concerning humans' perception of spaces and emotional reactions in real-time through using sensor technologies. This approach follows the field research of mass communication, media, sociologists, and criminologists regarding ethnically diverse residential communities. Based on the modelling of the Geographic Information System (GIS), mental maps, and spatial-statistical, Sorin Matei et al, (2001, p. 431) presents an approach for mapping the feelings of fear in the city of Los Angeles. They also produced the "emotional maps" based on the model of "*a dimensional Virtual Reality Modelling Language*." It strives at granting a better understanding of the environment.

Urban Sensing (People as Sensors): The Fourth Method:

This method focuses on the emotional cartography & bio-mapping. By using the internet (e.g. Wikimedia, OpenStreetMap, and Google Earth) and by using the developed mobile technologies or communication techniques. Urban Emotions is a trans-disciplinary, an interdisciplinary, human-centred approach that treats "people as sensors" to involve them in planning processes (Choudhury, Pradhan, & Kar, 2016, p. 27). It is the psycho-physiological approach to measure the arousal of people, depends on the data about the change in the body reactions (physiology) of people. It refers to gathering of the bio-sensor, biostatistical and physiological data (the startle reflex of the body, reflex of the neck or the eye) (Zeile, et al., 2015, p. 907). It is used to extract contextual emotion information from the data of technical sensor

and human sensor. The results can be used in urban planning for decision support and the evaluation of ongoing planning processes. It depends on the ground-truthing emotion measurements through people as sensors, the type of emotion, and the cause why an emotion occurred, the context of the emotion (Resch, Summa, Sagl, & Zeile, 2015, pp. 204-205). Most current sensors see a person's "additional heart rate", skin conductance, body temperature—the latter two variables are counter rotating, i.e., in case an emotional spike occurs, skin conductance rises, and body temperature drops because the person emits cold sweat (Resch, Summa, Sagl, & Zeile, 2015, p. 204). This method provides feedback to real-world processes in urban management and planning that will help to detect previously unseen urban patterns (Resch, Summa, Sagl, & Zeile, 2015, p. 210). In 2008, Christian Nold was the first one who combined emotional data with the help of physiological parameters (skin resistance level) and GPS in his project Bio mapping/emotion mapping. He used the wearer's Galvanic Skin Response (GSR) records, which is an indicator of emotional arousal in conjunction with their geographical location. It explores the implications of visualising intimate biometric data and emotional experiences using technology. Notably, the "use of the internet in combination with georeferenced data and ubiquitous mobile computers can be described as the next evolutionary step of the Internet, the 'Geoweb' or 'Web 3.0'" (Bergner, Exner, Memmel, Raslan, Talal, Taha, & Zeile, 2013, p. 409). This method uses some psychological studies that discuss the causes of changing the inner feelings of people. The documenting the feeling comes from the influence of the body (emotion or conscience). Scholars were able to measure these effects using some modern techniques that combining methods of "the spatial planning," "geo-informatics/ GIScience/ geographical information systems (GIS)," "computer linguistics," "sensor technology methods," and "an emotion-tracking technique." (Li, Hijazi, Reinhard Koenig, Zhong, & Schmitt, 2016). The process merges the concepts of semantic, geographic, temporal distance, semantic, geographic, and temporal clusters (Resch, Summa, Sagl, & Zeile, 2015, p. 210). The tool of this approach is as follow: measurements from calibrated bio-sensors, subjective observations. The general Urban Emotions concept includes four steps in two groups (Figure 4) (Resch, Summa, Sagl, & Zeile, 2015, p. 201).

Figure 4. The two groups and the four steps in each concept of urban emotion. Source: the authors



Crowdsourcing Approach Cognitive Appraisal: The Fifth Method:

This approach uses the cognitive assessment of the internet technologies and GIS. Alexey Tarasov et al. (2010) describe the crowdsourcing approach has tasks outsourced to a large group of non-expert individuals. This approach focuses on the context of labelling corpora, which is used in a field of machine translation, computer vision, and sentiment analysis. It is used to collect people's affective responses towards space. This approach depends on the investigation and collection of information by individuals who are not connected to each other, through the use of virtual reality sites on the Internet and reality. Assessing essential qualities of urban space with emotional and visual data based on the model '...for measuring emotions is based on the concept of cognitive appraisal, which categorises a relatively complex set of secondary emotions' (Li, Hijazi, Reinhard Koenig, Zhong, & Schmitt, 2016, p. 3).

The Internet of Things (IoT): The Sixth Method

This method is a novel paradigm that is rapidly gaining ground in the scenario of the modern wireless telecommunications. The basic idea of this concept is the pervasive presence around us of a variety of things or objects (e.g. Radio-Frequency Identification (RFID) tags, sensors, actuators, mobile phones, etc.). Through a unique addressing scheme, the method can interact with each other and cooperate with their neighbours to reach common goals (Atzori, Iera, & Morabito, 2010, p. 2787) (Giusto, Iera, Morabito, & Atzori, 2010). Besides, the method is considered an ecosystem powered by sensors and microchips, which enables connection and communication among real-world objects, environments, software, and people. This network of things, sensors and devices are capturing and exchanging enormous amounts of data and fueling the Big Data movement. And, "[t]he functionality of the IoT depends upon four fundamental steps: data acquisition, information extraction, knowledge extraction, and action-taking (L'Heureux, Grolinger, Higashino, & Capretz, 2017, p. 74). A smart environment is one of the relevance of this approach; it is "that making its employment" easy and comfortable things to the intelligence of "contained objects to be an office, a home, an industrial plant, or a leisure environment." It creates a) comfortable homes and offices. Furthermore, this method is used in the industrial plants, a smart museum and gym (Atzori, Iera, & Morabito, 2010, pp. 2795-2796).

CONCLUSION

This chapter referred to one of the most scientific aspects of the progress that may be the specialists in the fields of urban planning and design made in last few decades. This progress has been very clear on the track of developmental planning and design methods from the beginning of the sixties until now. Nowadays, it is reflected in the methods of testing in urban studies and research regarding its suitability, while also considering the circumstances of this technological age. Thus, various advanced ways emerged over the last three decades in the field of urban planning and design, as it has aimed to achieve accurate findings more than before. As a result, many modern trends in planning and design turned towards building the future cities based on the concepts of smart growth, the smart city, and the eco-city. All of them have been characterized as "intelligent" by the use of digital-based technology; they depend on intelligent buildings, smart urban citizens, smart environments, smart lifestyles, smart economies, and smart governance, as well as a smart integrated supply system for infrastructure. With the building of

these types of cities, digital-based technology has also emerged in smart energy, smart utilities, smart and integrated transportation, mobility system, and smart parking.

Two directions were considered in an example of the conventional and advanced research methods in this work. The first focused on how the urban planners and designers employ visual aesthetics and perceptual dimensions. The second direction deals with the new methods which employ the affective/emotional dimension by using digital tools. The review gave the following results:

- Two methods are included in the first set. The first is the contribution of Kevin Lynch (1960). He used the wayfinding as a method which revolves around how to guide people within the built environment, depending on the visuospatial representations of the city. Afterwards, specialists used the term of legibility based on the wayfinding method, such as: Jerry Weisman (1981); GD Weisman (1987); John Peponis, Craig Zimring, and Yoon Kyung Choi (1990); Paul Arthur and Romedi Passin (1992); Susan Hunter (2010). The second method presents a theory to study spatial interrelationships between people's interactions, called space syntax, by Bill Hillier and Julienne Hanson. It started as a concept in the early 1970s and became an approach in 1967, when it focused on the perceptual dimension of urban spaces. In 2008, it became a methodology and a tool. These methods are still developing to date.
- The second set of research focuses on the affective response approach, particularly humans' perception of spaces and emotional reactions. It started in the 2000s and includes human sensory assessment urban sensing, crowdsourcing, and the Internet of Things. These methods focus on emotional cartography and bio-mapping, the data of technical sensor and human sensor, semantic, geographic, and temporal distance, semantic, geographic, and temporal clusters, sentiment analysis, the concept of cognitive appraisal, and modern wireless telecommunications.

Nowadays, various tools depend on several digital technology and visualization methods and techniques:

- Integrating dynamic urban development simulation models and 3D visual urban simulation in visual impact analysis includes scale models, photo simulation, computer-aided design graphics, and video simulation. For instance, it has been used in urban planning and design disciplines to follow the public participation of historic preservation including visual impact analysis.
- Spatial planning, geoinformatics, geographical information systems, computer linguistics, sensor technology methods, and an emotion-tracking technique.
- The Internet web: Wikimedia, OpenStreetMap, Google Earth, and Web-based visualization technologies.
- Developed mobile technologies or communication techniques.
- Radio-frequency identification tags, sensors, actuators, and mobile phones.

These innovative methods and more should be considered in future research concerning the outgrowths of digital technology in the field of architecture in rehabilitation of traditional and historic cities, as well as in the process of design thinking and the planning of new cities. Prominently, it should consider not only physical dimensions but also non-physical aspects to establish future cities and sustainable and smart cities.

FUTURE RESEARCH DIRECTIONS

The current study confirmed that each research problem requires a specific method to extract the results. Nowadays, various new methods and techniques can be used in the field of urban planning and design such as (1) Isovist analysis 1979, (2) Visual Graph Analysis, or VGA 2001, (3) Vision-guided Agent-Based Modelling 2002, (4) Body Mass Index (BMI) 2002, (5) Segment-angular spatial network analysis 2005, (6) Place Syntax, a combined accessibility and the urban environment elements analysis tool 2008, (7) Origin-destination weighted network analysis to advance the methods of spatial accessibility analysis 2012, (8) Topological, angular, and metric radii of space syntax analysis (DepthmapX) 2011-2015, (9) GIS-linked tools and software to generate and use the space syntax analysis within a GIS environment 2015, (10) Axial integration map, (11) The multi-criteria analysis (MCA) 2016, (12) Wearable Technologies: concepts, methodologies, tools, and applications (2018).

These innovative methods and more should be considered in future research concerning the outgrowths of digital technology in the field of architecture in the area of rehabilitation of traditional and historic cities, as well as in the process of design thinking and the planning of new cities. Prominently, it should consider not only physical dimensions but also non-physical aspects to establish future cities and sustainable and smart cities.

REFERENCES

- Abusaada, H., & Elshater, A. (2019a). Internet of things interpreting a Myth: Internally displaced persons transform cities of hardship into chaos cities. In H. Abusaada, C. Voulguth, & A. Elshater (Eds.), *Handbook of research on digital research methods and architectural tools in urban planning and design* (pp. 47–81). IGI Global., doi:10.4018/978-1-5225-9238-9.ch004
- Abusaada, H., & Elshater, A. (2019b). A crowdsourcing approach in urban design: A bibliographic review of cities of singularity. In H. Abusaada, C. Voulguth, & A. Elshater (Eds.), *Handbook of Research on Digital Research Methods and Architectural Tools in Urban Planning and Design* (pp. 1–24). IGI Global., doi:10.4018/978-1-5225-9238-9.ch001
- Abusaada, H., Vellguth, C., & Elshater, A. (Eds.). (2019). *Handbook of research on digital research methods and architectural tools in urban planning and design*. IGI Global., doi:10.4018/978-1-5225-9238-9
- Arthur, P., & Passini, R. (1992). *Wayfinding: people, signs, and architecture*. McGraw-Hill Book Co.
- Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787–2805. doi:10.1016/j.comnet.2010.05.010
- Benedikt, M. (1979). To take hold of space: Isovists and isovist fields. *Environment & Planning*, 6(1), 47–65. doi:10.1068/b060047
- Bergner, B. S., Exner, J.-P., Memmel, M., Raslan, R., Talal, M., Taha, D., & Zeile, P. (2013). Human Sensory Assessment Linked with Geo and Mobile-Data Processing Methods in Urban Planning Exemplified on Different Cultures in Germany and Egypt. In *18th International Conference on Urban Planning, Regional Development and Information Society REAL CORP Conference*, (pp. 407–417). Rome, Italy: Academic Press.

Carmona, M., Heath, T., Oc, T., & Tiesdell, S. (2003). *Public Places-Urban Spaces, the Dimensions of Urban Design*. Architectural Press.

Choudhury, S., Pradhan, M. P., & Kar, S. K. (2016). A Survey on Determining Urban Emotions using Geo-Data Classification: A Case Study around Majitar, East District, Sikkim. *International Journal of Computers and Applications*, 135(2), 26–29. doi:10.5120/ijca2016908330

Dalton, R. C., & Bafna, S. (2003). The syntactical image of the city: a reciprocal definition of spatial elements and spatial syntaxes. *4th International Space Syntax Symposium*, (pp. (59)1-(59)21). London: Academic Press.

Giusto, D., Iera, A., Morabito, G., & Atzori, L. (Eds.). (2010). *The Internet of Things*. Springer. doi:10.1007/978-1-4419-1674-7

Haq, S., & Berhie, G. (2017). Space syntax investigation of Lubbock, a grid-like American city and some insights into isotropic layouts. *Journal of Urban Design*, 5–22. doi:10.1080/13574809.2017.1369872

Hillier, B. (1996). *Space is the Machine*. Cambridge University Press.

Hillier, B., & Hanson, J. (1984). *The social logic of space*. Cambridge University Press. doi:10.1017/CBO9780511597237

Hillier, B., & Hanson, J. (1997). The reasoning art: or, The Need for an Analytical Theory of Architecture. *Space Syntax: First International Symposium*, 1.

Hunter, S. (2010). *Design Resources: Spatial Orientation, Environmental Perception and Wayfinding*. University of Buffalo.

Jabbari, M., Fonseca, F., & Ramos, R. (2017). Combining multi-criteria and space syntax analysis to assess a pedestrian network: The case of Oporto. *Journal of Urban Design*, 23(1), 23–41. doi:10.1080/13574809.2017.1343087

Karimi, K. (2017). Space syntax: Consolidation and transformation of an urban research field. *Journal of Urban Design*, 23(1), 1–4. doi:10.1080/13574809.2018.1403177

Klettner, S., Huang, H., Schmidt, M., & Gartner, G. (2013). Crowdsourcing Affective Responses to Space. *Kartographische Nachrichten*, 2, 66–72.

Koenig, R., & Martin Bielik, S. S. (2018). System Dynamics for Modeling Metabolism Mechanisms for Urban Planning. In *Symposium on Simulation for Architecture and Urban Design (SimAUD)*. Delft, Netherlands: Society for Modeling & Simulation International (SCS).

L'Heureux, A., Grolinger, K., Higashino, W. A., & Capretz, M. A. (2017). A Gamification Framework for Sensor Data Analytics. In *IEEE International Congress on Internet of Things (ICIOT)* (pp. 74-81). Honolulu, HI: IEEE. 10.1109/IEEE.ICIOT.2017.18

Li, X., Hijazi, I., Reinhard Koenig, Z. L., Zhong, C., & Schmitt, G. (2016). Assessing Essential Qualities of Urban Space with Emotional and Visual Data Based on GIS Technique. *International Journal of Geo-Information*, 5(11), 1–18. doi:10.3390/ijgi5110218

Lynch, K. (1960). *The Image of the City*. The MIT Press.

- Matei, S., Ball-Rokeach, S. J., & Qiu, J. L. (2001). Fear and Misperception of Los Angeles Urban Space: A Spatial-Statistical Study of Communication-Shaped Mental Maps. *Communication Research*, 28(4), 429–463. doi:10.1177/009365001028004004
- Mehaffy, M. W. (2008). Generative methods in urban design: A progress assessment. *Journal of Urbanism*, 1(1), 57–75. doi:10.1080/17549170801903678
- Morello, E., & Ratti, C. (2009). A digital image of the city: 3D isovists in Lynch's urban analysis. *Environment and Planning. B, Planning & Design*, 36(5), 837–853. doi:10.1068/b34144t
- Peponis, J., Zimring, C., & Kuang Choi, Y. (1990). Finding the Building in Wayfinding. *Environment and Behavior*, 22(5), 555–590. doi:10.1177/0013916590225001
- Resch, B. (2013). People as Sensors and Collective Sensing Contextual Observations Complementing Geo-Sensor Network Measurements. In J. Krisp (Ed.), *Advances in Location-Based Services* (pp. 391–406). Springer International Publishing. doi:10.1007/978-3-642-34203-5_22
- Resch, B., Summa, A., Sagl, G., & Zeile, P. (2015). Urban Emotions—Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors and Crowdsourced Data. In G. Gartner & H. Huang (Eds.), *Progress in Location-Based Services 2016. Lecture Notes in Geoinformation and Cartography* (pp. 199–212). Springer. doi:10.1007/978-3-319-11879-6_14
- Sagl, G., Resch, B., & Blaschke, T. (2015). Contextual Sensing: Integrating Contextual Information with Human and Technical Geo-Sensor Information for Smart Cities. *Open Access Sensors*, 15, 17013–17035. doi:10.3390/150717013
- Tarasov, A., Cullen, C., & Delany, S. J. (2010). Using crowdsourcing for labelling emotional speech assets. *W3C workshop on Emotion ML*. doi:10.21427/D7RS4G
- Weisman, G. (1987). Improving wayfinding and architectural legibility in housing for the elderly. In V. Regnier & J. Pynoo (Eds.), *Housing the aged: Design directives and policy considerations* (pp. 441–461). Elsevier Science Publishing.
- Weisman, J. (1981). Evaluating Architectural Legibility Way-Finding in the Built Environment. *Environment and Behavior*, 13(2), 189–204. doi:10.1177/0013916581132004
- Wineman, J. D., & Peponis, J. (2010). Constructing Spatial Meaning: Spatial Affordances in Museum Design. *Environment and Behavior*, 42(1), 86–109. doi:10.1177/0013916509335534
- Zeile, P., Resch, B., Dörrzapf, L., Exner, J.-P., Sagl, G., Summa, A., & Sudmanns, M. (2015). Urban Emotions – Tools of Integrating People's Perception into Urban Planning. In *From Vision to Reality for Vibrant Cities and Regions, 20th Int. Conference on Urban Planning and Regional Development in the Information Society*, (pp. 907–914). Ghent, Belgium: Academic Press.

ADDITIONAL READING

- Aletta, F., & Xiao, J. (2018). *Handbook of Research on Perception-Driven Approaches to Urban Assessment and Design*. IGI Global. doi:10.4018/978-1-5225-3637-6

Anderson, B., & Ash, J. (2015). Atmospheric Methods. In P. Vannini (Ed.), *Non-Representational Methodologies: Re-Envisioning Research* (pp. 34–51). Routledge.

Berta, M., Bottero, M., & Ferretti, V. (2016). A mixed methods approach for the integration of urban design and economic evaluation: Industrial heritage and urban regeneration in China. *Environment and Planning. B, Planning & Design*, 1–25. doi:10.1177/0265813516669139

Goodchild, M. (1992). Geographical information science. *International Journal of Geographical Information Systems*, 6(1), 31–45. doi:10.1080/02693799208901893

Groux, S. L., & Verschure, P. F. (2010). Emotional Responses to the Perceptual Dimensions of Timbre: A Pilot Study Using Physically Informed Sound Synthesis. In S. Ystad, M. Aramaki, R. Kronland-Martinet, & K. Jensen (Ed.), *Exploring Music Contents: Proceedings of the 7th International Symposium on Computer Music Modeling*. Málaga, Spain.

Koenig, R., & Martin Bielik, S. S. (2018). System Dynamics for Modeling Metabolism Mechanisms for Urban Planning. *Symposium on Simulation for Architecture and Urban Design (SimAUD)*. Delft, Netherlands: Society for Modeling & Simulation International (SCS).

Mao, G., & Fidan, B. (2009). *Localization Algorithms and Strategies for Wireless Sensor Networks: Monitoring and Surveillance Techniques for Target Tracking*. IGI Global., doi:10.4018/978-1-60566-396-8

Nam, T., & Pardo, T. A. (2011). Conceptualizing Smart City with Dimensions of Technology, People, and Institutions. *The Proceedings of the 12th Annual International Conference on Digital Government Research* (pp. 282–291). USA.: College Park, MD.

Silva, C. N. (2018). *New Approaches, Methods, and Tools in Urban E-Planning*. IGI-Global., doi:10.4018/978-1-5225-5999-3

Wang, X., & Hofe, R. v. (2007). *Research Methods in Urban and Regional Planning*. Tsinghua University Press, Beijing and Springer-Verlag GmbH. doi:10.1007/978-3-540-49658-8

KEY TERMS AND DEFINITIONS

Pragmatic Method: Depends on the philosophy that inspires people to attend for a way and do the things by the best to achieve their aspired intentions and ends.

Research Method: A discipline deals with the principles and techniques of scientific interrogation, includes three approaches: quantitative, qualitative, and mixed methods.

Spatial Analysis: A research paradigm to explain the patterns of human behavior and their spatial expression through analyzing the events of a location or place, by focuses on two ways: buffering and overlaying.

Urban Emotion: An interdisciplinary and trans-disciplinary psycho-physiological approach and method; it provides feedback to real-world. People are the sensors focusing on the data about the change in body reactions. UE is to extract contextual emotion information from the data of “technical and human sensor” for decisions support and the evaluation of ongoing planning and design processes.

Urban Form: Consequences of two parts of the built environment (masses and spaces), which emerges in three-dimensional configuration of the physical characteristics related different scales. It responds to

the total societal impacts of the built environment: functional, socio-cultural, economic, behavioural, temporal, political, environmental, and technological dimensions.

Urban Sensing: Investigate the dynamics of the city considering people as sensors related the social networks in the community. It focuses on urban computing to handle the data through ubiquitous computing techniques.

Wayfinding: A strategy, approach, quantitative analytical method, and methodology to explore the spatial system and the configuration of the layout of the built environment, as well as it examines the effects of this configuration on the organization of the functions of the situation and the relations with the entire system.

Reengineering Legacy Systems Towards New Technologies

Djelloul Bouchiha

Center University of Naama, Algeria & EEDIS Lab. UDL-SBA, Algeria

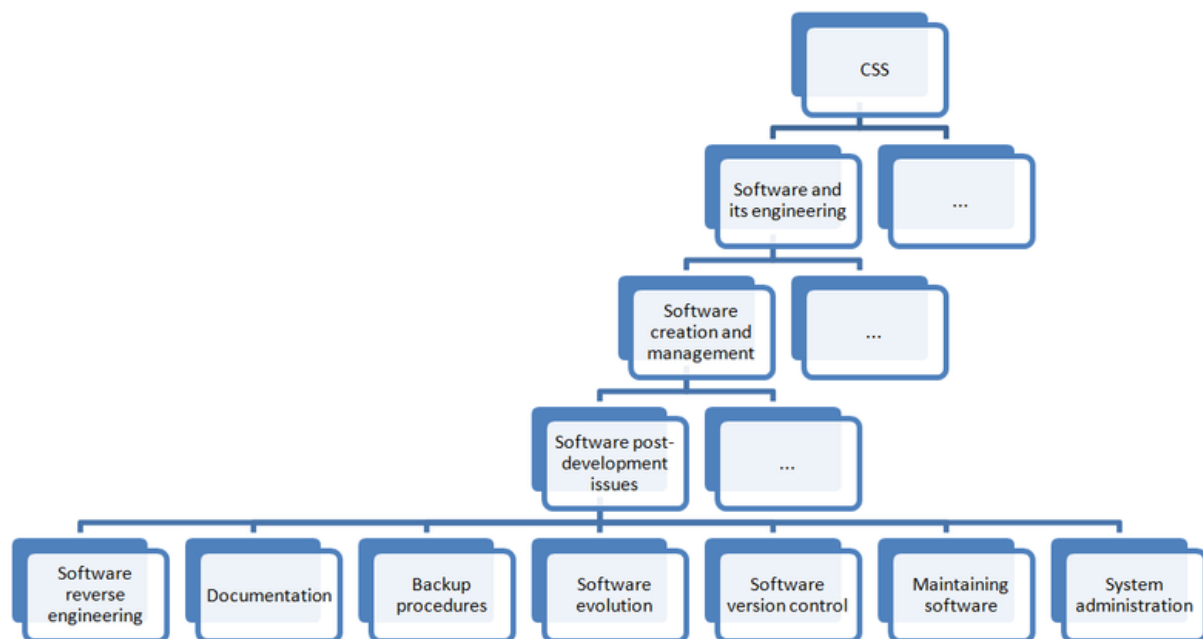
INTRODUCTION

Software reengineering is an important area of the software engineering. The quest to maintain and understand operational legacy systems has always been a challenge for software practitioners. This chapter presents a compilation of notions and techniques covering major areas namely, reverse engineering, program understanding, software maintenance, migration and evolving. Our objective is not to create new terms, but to introduce the terms already in use with different perspectives.

Before starting with the reengineering of legacy systems towards new technologies, we must know: in which research field are we working? Which research community do we belong to? What research issue are we dealing with?

To answer all these questions, we use the ACM Computing Classification System (CCS). It has been built as a hierarchical ontology that can be used in semantic Web applications. It relies on a semantic terminology as the main source of concepts and categories that present the state of the art of the computing discipline. The 2012 ACM CCS (ACM-CCS, 2012) starts with thirteen computing fields. Browsing this hierarchy to attempt our goal gives us the following tree.

Figure 1. Reengineering issues in the ACM CSS classification



DOI: 10.4018/978-1-7998-3479-3.ch084

Leaves of the presented tree in Figure 1 are the sub-issues in relation to the reengineering problem, according to the ACM CSS classification.

The following sections give more details about the reengineering process, notably prerequisites, model, problem statement, taxonomy, techniques and future trends. We also list some specialized conferences and journals. Finally, we give conclusion and advices to the researchers belonging to the software engineering community.

BACKGROUND

This chapter is titled “Reengineering Legacy Systems towards new Technologies”. Here, the term “System” is used to say “software system” not the hardware one. “Legacy systems” are systems currently in use, and have been often left to run for a long time without too many significant changes. “New technologies” are the latest (current) technologies.

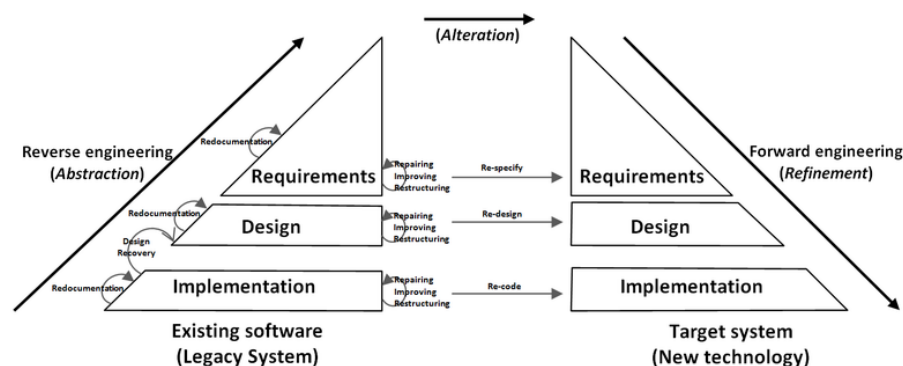
Reengineering is discussed in this chapter as sub-issue of Software Engineering which is described as the application of engineering techniques to produce software that runs properly and is constructed in an efficient manner. The produced software should be efficient, reliable, modifiable, testable, reusable, maintainable, portable, interoperable and correct. Software engineering process consists of a set of activities, known as the software lifecycle. Principal activities are: requirements analysis, design, implementing and testing the software. Requirements analysis phase consists in determining precisely what the functionality of the system will be. The design phase consists in taking the requirements and devising a representation to allow the requirements to be translated into source code. The implementing activity is the most familiar to students that consists in writing the program source code. Software testing consists in checking the software before it is judged ready to be released to its customer.

Reengineering is the study and analysis of an existing system for purposes of understanding, maintenance, or migration towards new technologies (Chikofsky & Cross, 1990). Based on the software lifecycle described above, the next section presents a general model for the software reengineering process.

REENGINEERING MODEL

We propose, in Figure 2, a general model for software reengineering that depicts the sub-processes for all levels of the reengineering process.

Figure 2. General model for software reengineering



Reengineering starts with the source code of an existing software (legacy system), and finishes with the source code of a target system. This is done through three sub-processes: **(1) Abstraction** is a gradual increase in the abstract level of system. This upward movement is called Reverse engineering. **(2) Alteration** is changing the system representation at the same level of abstraction. According to the abstraction level, it can be *recoding* (at the implementation level, to translate the existing code to a target language), *redesigning* (at the design level) or *respecifying* (at the requirement level). **(3) Refinement** is a gradual decrease in the abstract level of system. This is termed forward engineering.

In fact, the reengineering process involves maintenance or migration, according to the used activities:

Example 1: going from the legacy system (*implementation*) to the *design* abstraction level where we correct some errors (*repairing*), and then we return to the implementation with the same language: this is maintenance process.

Example 2: going from the legacy system (*implementation*) to the *design* abstraction level, then we alter the design representation (*redesign*), and finally we generate a new code with a new language: this is a migration process.

PROBLEM STATEMENT AND PROPOSED SOLUTION

In this section, we answer the question: Why do we need software reengineering?

So the reengineering process aims at migrating, maintaining or understanding legacy systems (Alam & Padenga, 2010):

1. *Migration issue:* the number of legacy systems currently in use is very high. These systems cannot be abandoned because they are applications in which companies have already invested considerable time and money.

Existing software may be based on hardware platforms, operating systems, or languages that might have become obsolete. On the other hand, the computer industry is growing at a fast rate; new hardware and software systems with new features, quickly outdating legacy systems and thus replacing them with new systems built from scratch can be disruptive. Therefore, it will be more rational and economic to reuse and adapt legacy systems to these new technologies.

2. *Maintenance issue:* over time and with multiple changes in the existing software system, it will be of low quality. As maintenance and changes of the software continue, its reliability steadily decreases to the point of being unacceptable.

Legacy systems lack a good design structure and code organization, making changes to the software difficult and costly. Enhancements of the functionality of the existing systems may also be needed.

3. *Understanding issue:* the difficult part lies in understanding of the existing system. In most cases the requirements, design and code documentation will be no longer available, or will be out of date, so it is really not clear what functions are to be moved. In most cases the original developers of the legacy systems are not available to verify or explain the system; the only available source is the current software source code.

For these reasons, companies with working software that meets their needs might need to migrate to new hardware platform, operating system or language. So, a reengineering process is required to migrate our existing systems into more advanced platforms that are easier to maintain.

Reengineering process consists in taking existing software that have become expensive to maintain or whose system architecture or implementation are now obsolete, and redoing it with up to date technology. It allows taking legacy systems, and generating a new target system that preserves the required functionality, while applying new technologies. It allows also increasing reliability, improving maintainability, reducing the cost of maintenance, and preparing for functional enhancement.

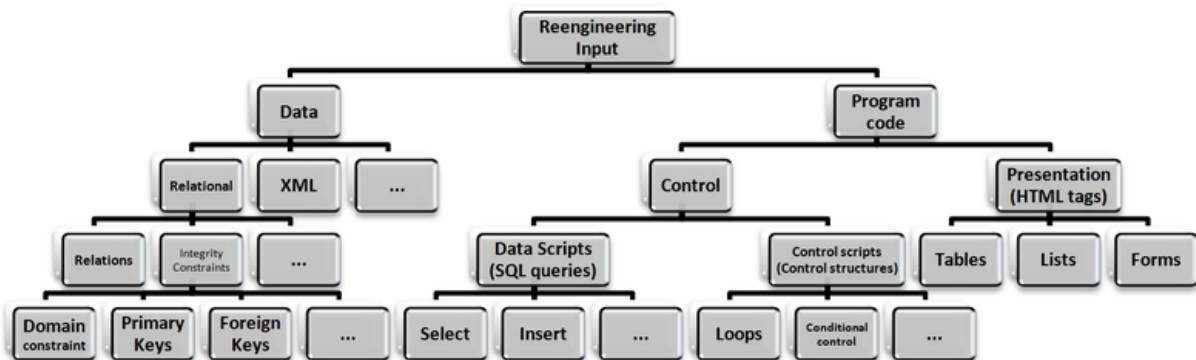
The reengineering process is mainly characterized by its inputs. This is why we propose in the next section the inputs taxonomy.

REENGINEERING INPUTS TAXONOMY

The procedure for classifying reengineering approaches of the legacy systems is generally based on their inputs. Inputs are resources that can be used to generate data and control models, which in turn can be used in the subsequent steps of the reengineering process. This process starts from the implementation level where there exists often data and program code.

After going through various literatures in the reengineering field, we were able to propose reengineering inputs taxonomy as depicted in Figure 3.

Figure 3. Reengineering taxonomy according to the system Inputs



At the first level, the reengineering process can be classified into two categories, according to the system input:

1. *Data-driven Reengineering process*: relational, XML... data can be analyzed using Application Interfaces (API), in order to recover data models (class diagram, entity-relationship model ...) at the reverse engineering stage. In the literature, recovering data models is often called “database reverse engineering”.

For XML files, the DTD (Document Type Definition) and XML schema play a very important role in the reverse engineering process, as they provide a very important resource from which a data model

can be constructed. For a relational DB, it is the relations (tables) and integrity constraints (Domain constraint, primary keys, foreign keys ...) that play this primordial role.

Mapping rules from entity-relationship to relational model (Bouchrika, 2014), or from class diagram to relational model (Ambler, 2000), can be used inversely to recover data models.

2. *Program-driven Reengineering process*: according to their objective, code fragments can be classified into two sub-categories:
 - a) Control code: program understanding techniques are used to analyze
 - Data scripts (SQL Queries): allow recovering data models from SQL queries (Select, Insert...) imbedded in the source code.
 - Control scripts (Control structures): allow recovering control models (activity diagram, system architecture, workflow models...) from iterative control (loops), conditional control (if-then-else, case)...
 - b) Presentation code: HTML tables, lists and forms, can be analyzed with API, in order to recover data models (class diagram, entity-relationship model, schema, DOM...).

We can't leave this section without providing descriptions of the terms used above:

Application Interface (API): an applicative layer used as intermediate between the application and resources. So, it aims at exploring data. For example JDOM Java API is used to explore HTML and XML files.

Model, Diagram & Schema: in literature there is ambiguity between these three terms.

- Model: is the result of an abstraction of a part of the real world; modeling is the abstraction process.
- Diagram: is a graphical representation of the model, for example class diagram in UML.
- Schema: the term schema is often related to another term. For example: XML schema, Database schema, electrical schematic, etc. In databases, the term "schema" refers to the organization of the database. So the schema describes the structure of the database. It indicates which tables (relations) constitute the database, as well as the attributes included in each table.

Why HTML tables, lists and forms are used as reengineering Inputs? because they are the most used elements to represent structured data on the Web. Tables and lists are used as data output, and Forms as data input.

PROGRAM UNDERSTANDING TECHNIQUES

Program-driven reengineering process, described in the previous section, needs program understanding techniques. Müller (1996) defines *program understanding* (PU) or **program comprehension** as "the task of building mental models of the underlying software at various abstraction levels, ranging from models of the source code itself to ones of the underlying application domain for maintenance, evolution and reengineering purposes".

Many techniques have been developed to help the analyst to understand existing programs:

1. Pattern matching: allows looking for patterns in the program source code.
2. Variable dependency graph: is a graph that represents relations between the variables.
3. System dependency graph analysis: is a program representation used by program slicing.
4. Program slicing: is a decomposition technique that extracts from a program, statements that come into play in a particular computation.
5. Graphical visualization of program: some aspects of the program architecture can be represented as graphs.

Pattern matching:

Patterns are known as plans, cliches, motifs, molds, templates, string... Whatever the name, pattern aims to identify predefined structure in the program source code.

A Pattern Definition Language can be used to describe the patterns. For example, pattern of the Select SQL query can be defined as follows:

```
Select <attribute>*
from <table>*
where (<column> = <value>)*;
```

Tilley et al. distinguish between structural and behavioral patterns, depending on whether one is looking for code that has a specified syntactic structure, or searching for code components that share specific data-flow, control-flow or dynamic relationships (Tilley, Paul, and Smith, 1996).

The remaining program understanding techniques use, all of them, directly or indirectly, pattern matching to achieve their objectives.

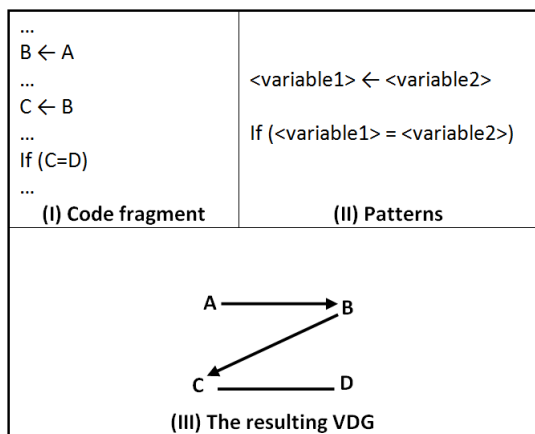
Variable dependency graph:

In the variable dependency graph (VDG), each variable of the program is represented as a vertex, while an arc (directed or not) represents a direct relation (assignment, comparison, etc.) between two variables.

To construct this graph, it is only needed to look for definite statement patterns in the program.

Figure 4 shows the analyzed fragment of the program in (I), the used patterns to recognize relations between variables in (II), and the resulting VDG in (III).

Figure 4. Example of variable dependency graph (VDG)



A path, between variable A and variable D in the graph, reflects the existence of a sequence of statements in the program such that the value of A is in relation with the value of D.

The interpretation of “the variable A being in relation with the variable D” can be one of the following: the two variables belong to the same domain or share the same values, the two variables denote the same real world object, there is a dependency between the two variables, etc.

System dependency graph:

The system dependency graph (SDG) is an extension of the program dependency graph (PDG).

A- Program dependency graph

While VDG represents dataflow, PDG represents the control flow.

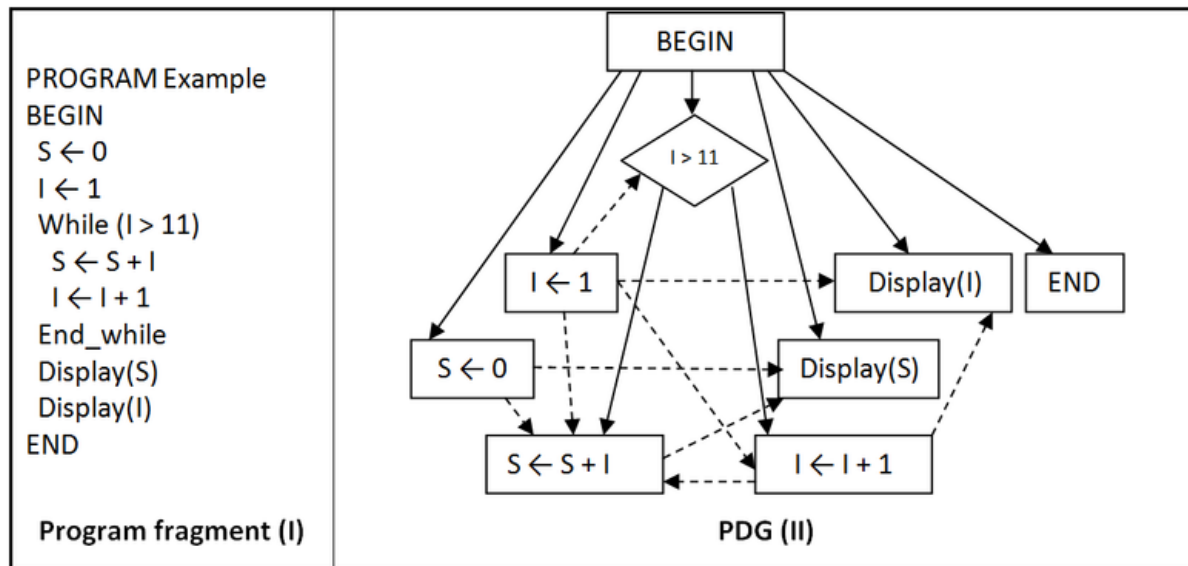
A PDG is a directed graph, where vertices correspond to statements and control predicates, and edges correspond to data and control dependencies.

A control dependence edge from vertex X1 to vertex X2 means that the execution of the statement X2 depends on the evaluation result of the predicate X1.

A data dependence edge from vertex X1 to vertex X2 means that the program’s computation might be changed if the relative order of the components represented by X1 and X2 were reversed.

Figure 5 shows a program fragment (I) and its corresponding program dependency graph (II). The solid arrows represent control dependence edges, and the dashed arrows represent data dependence edges.

Figure 5. Example of program dependency graph (PDG)



Note that control dependences connect always the entry vertex (BEGIN) with the procedure’s statements and predicates.

B- System dependency graph

Horwitz et al. introduced the term system dependency graph (SDG) for the dependency graphs that represent multi-procedure graphs (Horwitz et al., 1990).

A system dependency graph (SDG) includes a program dependency graph (PDG), which represents the system's main program, and procedure dependency graphs, which represent the system's auxiliary procedures, and some supplementary edges. These supplementary edges are of two sorts: (1) edges that represent direct dependencies between a procedure call site and the called procedure, and (2) edges that represent transitive dependencies due to procedure calls.

Figure 6. Example of system dependency graph (SDG)

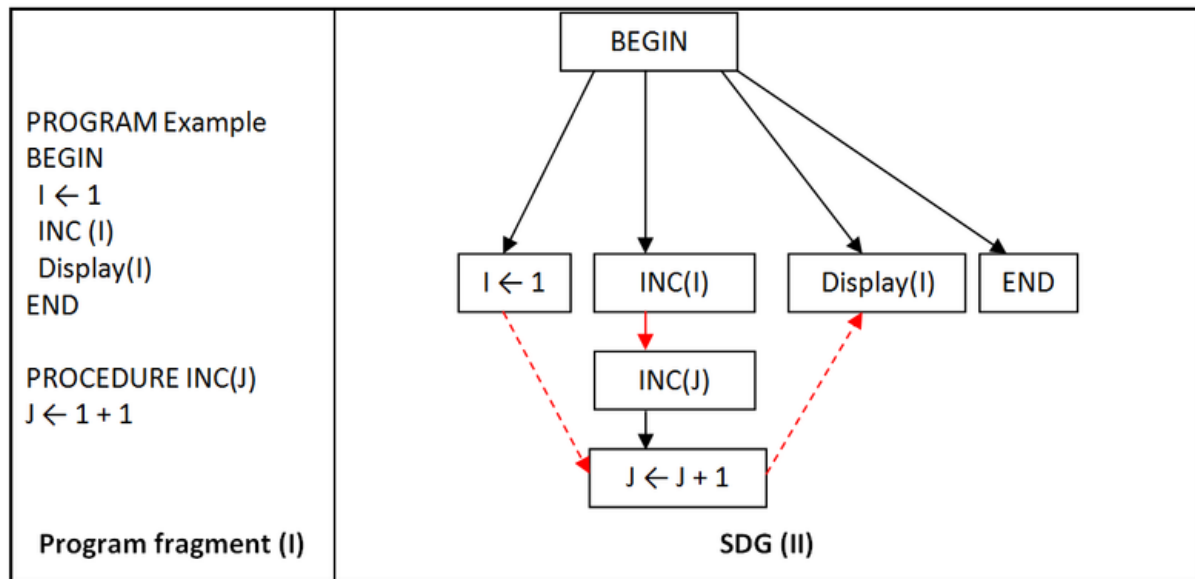


Figure 6 shows a program fragment (I) and its corresponding system dependency graph (II). Red arrows represent the supplementary edges. A solid red arrow is between the call site and the called procedure, and dashed arrows represent data dependence edges between program statements and procedure statements.

Program slicing:

Program slicing is a decomposition technique that extracts statements from a program (source code) that are relevant to a particular computation.

Informally, a slice answers the question “Which are the program statements that potentially affect the computation of variable X at the point A?”; so a slice of a program consists of all statements and predicates that might affect the value of X at the point A.

The task of computing program slices is called program slicing.

Program slicing uses the system dependency graph (SDG) to represent the program, and an algorithm (Horwitz, Reps and Binkley, 1990) to compute the slice.

Number of methods to compute slices have been proposed. A distinction is defined between static and dynamic slicing methods. The difference between static and dynamic slicing is that dynamic slicing admits a fixed input for a program, whereas static slicing does not make any assumption regarding the input.

Graphical visualization of the program:

Among graphs that can help understanding programs, Henrard in (Henrard, 2003) cites:

- Inter-program call graph: the vertices represent programs and the edges represent the call between programs.
- Intra-program call graph: represents the call relations between procedures in the program.
- Database usage graph: vertices are of two types (programs and the collections of entity-type), and edges link programs with the collections entity-type they use.
- JCL graph: represents the sequence of execution of the different programs with the files they use. Vertices are programs or files, and edges represent program sequences and file usage.
- Hyperlink graph: vertices represent documents, and edges represent the hyperlinks between these documents. This kind of graph can be used for the Web documents.

In the next section, we present a reengineering process based on models.

MODEL DRIVEN ENGINEERING (MDE)

One of the most important conceptual tools to implement a reengineering process is MDE. The model-driven engineering (MDE) is a generative process enabling to get, treat and produce information from other data derived from structured entities called models. MDE settings are mainly based on modeling, meta-modeling and transformation aspects. Another important MDE aspect is the conformity relationship. A model conforms to the meta-model as a program conforms to the programming language's grammar in which it is written. This means that a meta-model describes the different types of model elements and how they are arranged.

Therefore, the MDE is considering models and software as first class entities. The model represents any aspect of software, and it consists of a set of model elements.

Figure 7. A model-based transformation process according to MDE

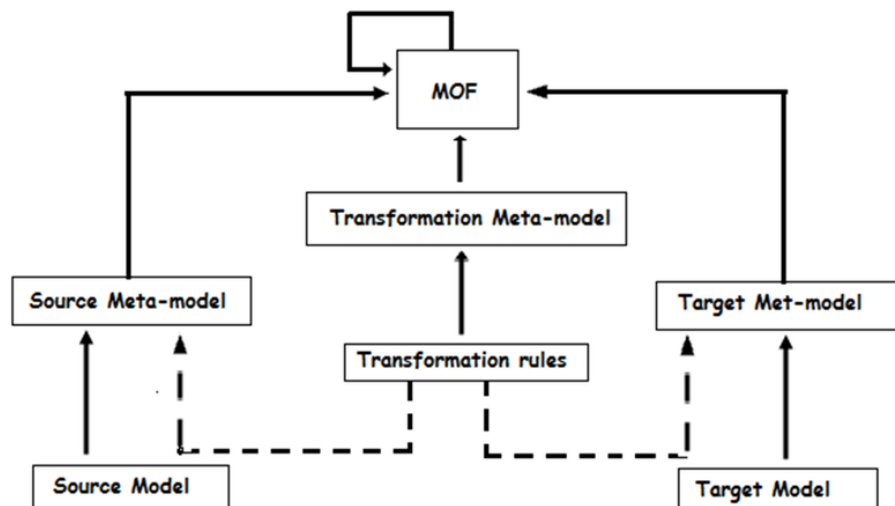


Figure 7 shows a model-based transformation process. It allows transforming any source model Ma that conforms to MMa (source meta-model), into a target model Mb that conforms to MMb (target meta-model). The transformation is driven by a set of rules expressed as transformation model, which itself conforms to a transformation meta-model (Bezivin, 2006).

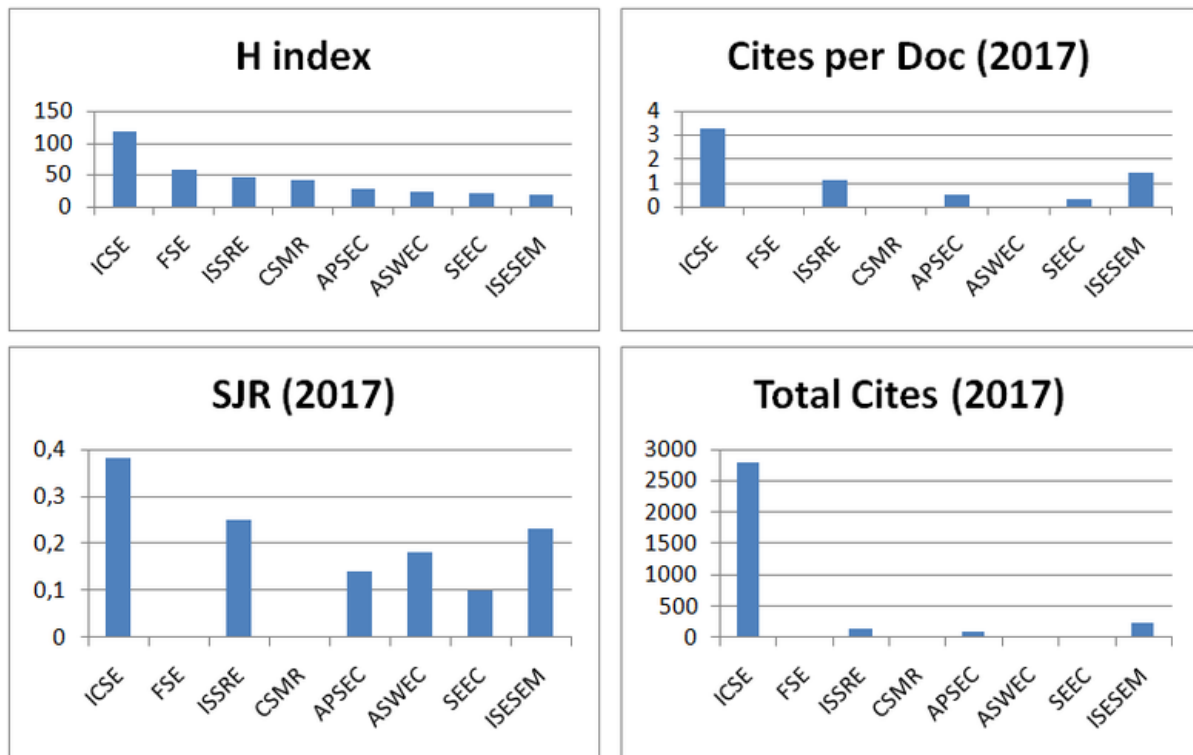
In the next section, we cite some specialized conferences and journals in the software engineering field.

SPECIALIZED CONFERENCES AND JOURNALS

In literature, there are many peer-reviewed journals and international conferences, where articles are published in the software engineering field. In peer-reviewed (refereed) journals or conferences, articles are reviewed by several experts in the field before they are published, in order to ensure their quality.

The following graph presents ranking information about some important Software Engineering conferences. This information has been summarized with help of SJR (Scimago Journal & Country Rank) (SJR, 2018).

Figure 8. Histograms of Software Engineering conferences ranking



“H Index” means that H papers have been published. Each one has been cited in other papers at least H times. “SJR” (SCImago Journal Rank) indicator indicates the average number of weighted citations of the conference proceedings. “Cites per Document” column indicates the average citations per document. “Total Cites” column corresponds to the number of citations in the selected year.

Figure 8 shows that **ICSE (International Conference on Software Engineering)** is the most important one. Since 1975, ICSE is a software engineering conference that provides a forum for researchers to propose and discuss recent innovations, trends, experiences and concerns in the field of software (re) engineering.

On its Web site (<http://www.icse-conferences.org/>) you can find all information about the ICSE conference series. You find also useful information for organizing the future ICSE conferences.

Table 1 explains more about the scope of the conferences presented in Figure 8.

Table 1. Conferences' scope

Conference	Aims and Scope
International Conference on Software Engineering (ICSE)	ICSE presents and discusses the most recent innovations, experiences, trends and concerns in the field of software engineering.
the ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE)	FSE presents and discusses the most recent innovations, trends, experiences, and challenges in the field of software engineering.
International Symposium on Software Reliability Engineering (ISSRE)	ISSRE covers reliability engineering for software. Reliability engineering is a sub-discipline of software engineering that emphasizes dependability in the lifecycle of a product.
Euromicro Conference on Software Maintenance and Reengineering (CSMR)	CSMR aims at uniting divergent groups of software maintenance professionals and academic software engineers via the addition of commercial suppliers.
Asia-Pacific Software Engineering Conference (APSEC)	APSEC aims to bring together researchers and practitioners from industry, academia, and government to encourage wider collaboration between academics and industries and to advance the state of the art in Software Engineering field.
the Australian Software Engineering Conference (ASWEC)	ASWEC attracts contributions that cover the whole spectrum of software engineering research and practice. It provides an opportunity for interaction between industry practitioners and software engineering researchers.
Software Engineering Education Conference (SEEC)	SEEC is looking for contributions that address challenges, innovations, and best practices in software engineering education.
International Symposium on Empirical Software Engineering and Measurement (ISESEM)	The objective of ISESEM is to provide a forum where researchers, practitioners, and educators can report and discuss the most recent research results, innovations, trends, experiences, and concerns in the field of empirical software engineering and metrics.

With the same principle, Figure 9 presents ranking information about the some specialized journals in the software engineering field, also with help of SJR (Scimago Journal & Country Rank) (SJR, 2018).

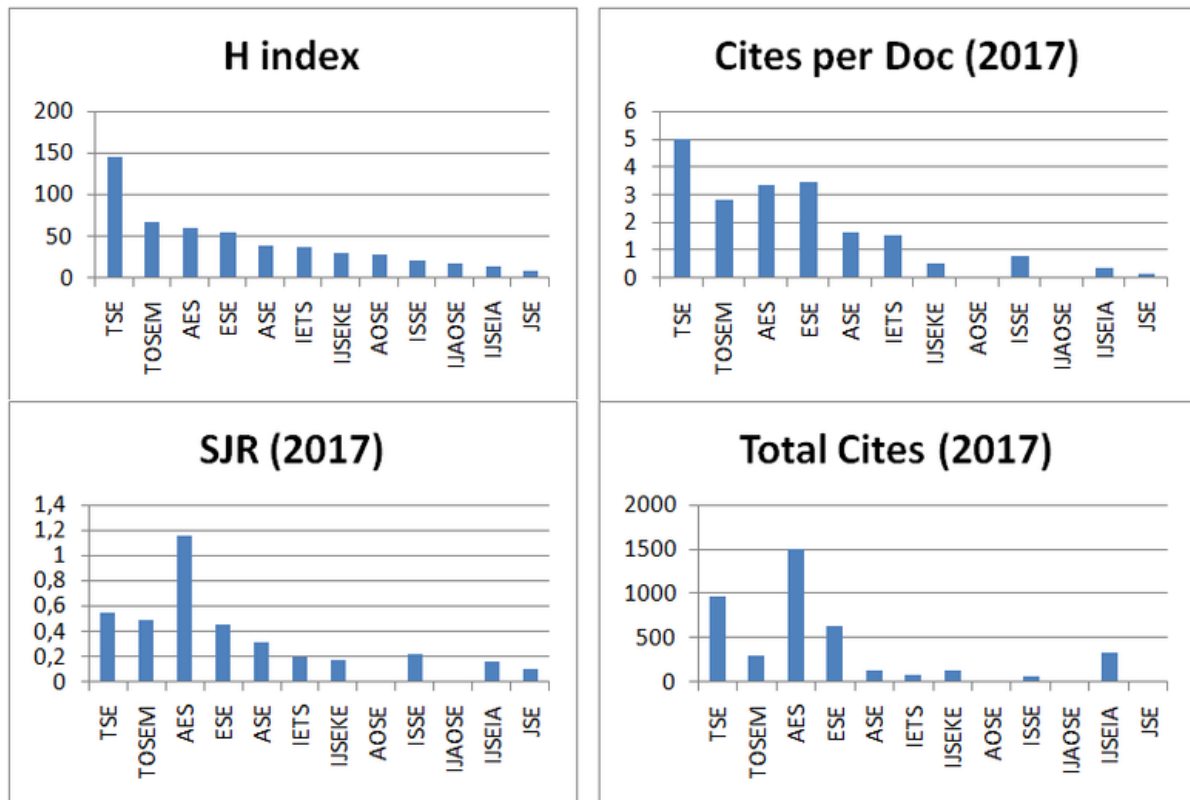
It is clear that the **IEEE Transactions on Software Engineering (TSE) journal** has the best ranking in terms of H index and cites per document. However, **Advances in Engineering Software (AES) journal** is a powerful competitor, when we conceder SJR and Total cites factors.

IEEE Transactions on Software Engineering journal focuses on well-defined theoretical results and experimental studies that have potential impact on the development, analysis and management of software. Its scope ranges from developing principles to applying those principles to specific environments.

The **Advances in Engineering Software journal** is interested in recent and projected advances in computer-based engineering techniques. It covers several fields, such as mechanical, aerospace, civil and environmental engineering, while concentrating on research and development leading to practical problem-solving.

Table 2 explains more about the scope of the journals presented in Figure 9.

Figure 9. Histograms of Software Engineering journals ranking



FUTURE RESEARCH DIRECTIONS

In the last ten years, some works focus on Web reverse engineering (Bouchiha, Malki & Mostefai, 2009a; Benslimane, Malki & Bouchiha, 2009). Web application is a software system that provides its functionality through the Web (Conallen, 1999). For their goal, most of these Web application reverse engineering approaches adapted techniques of database reverse engineering to generate conceptual schema.

In (Perez-Castillo, Garcia-Rodriguez, Caballero, Polo & Piattini, 2009; Bouchiha, Malki & Mostefai, 2009b; Niiyama, Chung, Chinn & Davolos, 2008; Bouchiha, Malki & Mostefai, 2008; Takonewa, Mamello & Paul, 2010), researchers aim at constructing Web Services from existing Web applications. Web Services are software components that provide services via a standardized interface. As a consequence, Web applications can be easily integrated to each other in a business-to-business schema to provide a more valuable service to users.

Actually, the trend is to build a technology stack to support a “Web of data”, which is the sort of data found in databases, instead of the “Web of documents”. The ultimate goal of the Web of data is to enable machines to do more useful works, and to develop systems that can support trusted interactions over the Web. The term “Semantic Web” refers to W3C’s vision of the Web of linked data (W3C, 2018a). Semantic Web technologies enable users to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data (Berners-Lee, 2006) are empowered by technologies such as RDF (Resource Description Framework) (W3C, 2018b). Several attempts have addressed the issue of reengineering legacy systems, notably Web applications, into Semantic Web (Lehmann et al. 2014;

Table 2. Journals' scope

Journal	Aims & Scope
IEEE Transactions on Software Engineering (TSE) (https://www.computer.org/web/tse)	TSE is interested in well-defined empirical studies and theoretical results that have potential impact on the analysis, construction, or management of software.
ACM Transactions on Software Engineering and Methodology (TOSEM) (https://tosem.acm.org/)	The scope of TOSEM includes models, methods, mechanisms, languages, and tools for the evaluation, elaboration, and evolution of products and processes all along the software lifecycle, from requirements specification to software maintenance.
Advances in Engineering Software (AES) (https://www.journals.elsevier.com/advances-in-engineering-software)	The scope of AES includes innovative computational strategies for large-scale engineering issues, analysis and simulation methods and systems, etc.
Empirical Software Engineering (ESE) (https://www.springer.com/computer/swe/journal/10664)	ESE aims at promoting the publication of industry-relevant research, since one of its aims is to address the significant gap between research and practice.
Automated Software Engineering (ASE) (https://link.springer.com/journal/10515)	ASE details research papers, tutorials, surveys and accounts of significant industrial experience in the foundations, techniques and tools of the software engineering technology.
IET Software (IETS) (https://digital-library.theiet.org/content/journals/iet-sen)	IETS publishes papers on all aspects of the software lifecycle, including design, development, implementation and maintenance. The focus of the journal is on the methods used to develop and maintain software, and their practical application.
International Journal of Software Engineering and Knowledge Engineering (IJSEKE) (https://www.worldscientific.com/worldscinet/ijseke)	IJSEKE covers areas relating to software engineering and knowledge engineering. The scope of the journal includes object-oriented systems, logic programming, rapid prototyping, and software and knowledge-ware maintenance.
Annals of Software Engineering (AOSE) (https://www.springer.com/computer/swe/journal/10480)	AOSE covers all aspects of software engineering. This includes Computer-Aided Software Engineering, Human Factors in Software Engineering, etc.
Innovations in Systems and Software Engineering (ISSE) (https://link.springer.com/journal/11334)	ISSE addresses problems and innovations in Software Engineering, Systems Integration, Software Development and other related fields.
International Journal of Agent-Oriented Software Engineering (IAOSE) (http://www.inderscience.com/jhome.php?jcode=IAOSE)	IAOSE fosters discussion on all software engineering aspects of the use of agent technology for the development of IT systems. This includes Software engineering methodologies for agents, MAS and agreement technologies, etc.
International Journal of Software Engineering and its Applications (IJSEIA) (http://www.sersc.org/journals/IJSEIA/)	IJSEIA aims at facilitating and supporting research related to software engineering and its applications. The scope of the journal includes Agent Technology, Aspect-Oriented and Feature Interaction, etc.
Journal of Software Engineering (JSE) (https://scialert.net/jhome.php?issn=1819-4311)	JSE is dedicated to publish original and significant results in all areas of software engineering. The scope of the journal includes methods and practices, object-oriented systems, rapid prototyping, etc.

Nagy et al. 2011; Mulwad, Finin & Joshi, 2013; Scharffe et al. 2012; Munoz, Hogan & Mileo, 2014; Cresci et al., 2014; Bouchiha, Malki, Alghamdi & Alnafjan, 2017).

The recent technological overhangs have focused on the miniaturization of communication devices, like Smartphone, in wireless networks. In this new hardware context, Ubiquitous Computing is a paradigm whose goal is to allow users to access data, and make them available anywhere and at any time. Pervasive Information Systems (PIS) constitute a category of Software Systems working in the Ubiquitous Computing environment. PIS cover a dynamic and complex environment instead of personal

computers only, perceive contextual information instead of simple user input, and support mobility instead of stationary services. Authors of (Hauser, Günther, Flath & Thiesse, 2017) pointed out the need to convert legacy systems to PIS. Our chapter suggests a set of reengineering techniques that can be used to address this issue.

CONCLUSION

Reengineering is the study and analysis of legacy system in order to understand, maintain or migrate it towards new technologies. In this chapter, definitions, techniques and even a list of specialized conferences and journals have been introduced; all of them concern the software reengineering research field. Illustrative models have been proposed and relevant questions have been answered to highlight the ways to authors who are interested in writing research papers or surveys about software reengineering.

We note that it is very difficult to write surveys about software engineering fields. So we advice authors of research papers to cite only works in relation with their ultimate goal (maintenance, migration, understanding...), and with the input and output of their proposed process (models, source code...). Evaluation, and specially comparison, is also difficult because of lack of common corpus, on which experiments can be done. Researchers have to choose their own corpus, which should be open access, and they can use some metrics, like those used in information retrieval (RI), notably Recall, Precision and F-measure. In the Evaluation step of a software engineering process, the expert intervention is always needed.

Finally, we hope that the readers, especially those belonging to the software engineering community, find this chapter useful as it provides an indicative overview of the concepts and tools in the area of software reengineering. The mastery of all these tools will provide significant incremental enhancement to software productivity.

REFERENCES

- W3C. (2018a). *Semantic Web*. Retrieved November 24, 2018, from <https://www.w3.org/standards/semanticweb/>
- W3C. (2018b). *Resource Description Framework (RDF)*. Retrieved November 24, 2018, from <https://www.w3.org/RDF/>
- ACM Computing Classification System (CCS). (2012). Retrieved from <https://www.acm.org/publications/class-2012>
- Alam, M. A., & Padenga, T. (2010). *Application Software Reengineering*. Dorling Kindersley (India) Pvt. Ltd.
- Ambler, S. W. (2000). *Mapping objects to relational databases*. Retrieved from <https://www.ibm.com/developerworks/library/ws-mapping-to-rdb/>
- Benslimane, S. M., Malki, M., & Bouchiha, D. (2009). Maintaining web application: An ontology-based reverse engineering approach. *International Journal of Web Information Systems*, 5(4), 495–517. doi:10.1108/17440080911006225

Berners-Lee, T. (2006). *Linked data - design issues*. Retrieved from <http://www.w3.org/DesignIssues/LinkedData.html>

Bezivin, J. (2006). Model driven engineering: An emerging technical space. In *Generative and Transformational Techniques in Software Engineering* (Vol. 4143, pp. 36–64). Springer.

Bouchiha, D., Malki, M., Alghamdi, A., & Alnafjan, K. (2017). HTAB2RDF: Mapping HTML tables to RDF triples. *Journal of COMPUTING AND INFORMATICS*, 36(6), 1467–1491. doi:10.4149/cai_2017_6_1467

Bouchiha, D., Malki, M., & Mostefai, A. (2008). A Framework for Reengineering Web Applications to Web Services. *Journal of Computational Science*, 7, 19–26.

Bouchiha, D., Malki, M., & Mostefai, A. (2009a). An Extended Approach for Recovering BPMN and WebML Models From Legacy Web Applications. *Journal of Computational Science*, 8(3), 40–48.

Bouchiha, D., Malki, M., & Mostefai, A. (2009b). Towards reengineering Web applications to Web services. *The International Arab Journal of Information Technology*, 6(4).

Bouchrika, I. (2014). *How to Convert ER Diagram to Relational Database*. Available at <http://www.learnadb.com/databases/how-to-convert-er-diagram-to-relational-database>

Chikofsky, E. J., & Cross, J. H. (1990). Reverse engineering and design recovery: A taxonomy. *IEEE Software*, 7(1), 13–17. doi:10.1109/52.43044

Conallen, J. (1999). *Building Web Applications with UML*. Object technology. Addison-Wesley Longman.

Cresci, S., DErrico, A., Gazze, D., Lo Duca, A., Marchetti, A., & Tesconi, M. (2014). Towards a DBpedia of Tourism: the case of Tourpedia. *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference (ISWC 2014)*.

Hauser, M., Günther, S. A., Flath, C. M., & Thiesse, F. (2017). Designing Pervasive Information Systems: A Fashion Retail Case Study. *Proceedings of the 38th International Conference on Information Systems (ICIS 2017)*.

Henrard, J. (2003). *Program understanding in database reverse engineering* (PhD Thesis). FUNDP-Namur Belgium.

Horwitz, S., Reps, T., & Binkley, D. (1990). Interprocedural slicing using dependence graphs. *ACM Transactions on Programming Languages and Systems*, 12(1), 26–60. doi:10.1145/77606.77608

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2014). Dbpedia a large scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 6(2), 167–195. doi:10.3233/SW-140134

Müller, H. (1996). Understanding Software Systems Using Reverse Engineering Technologies Research and Practice. *Proceedings of 18th International Conference on Software Engineering*.

Mulwad, V., Finin, T., & Joshi, A. (2013). Semantic Message Passing for Generating Linked Data from Tables. In *Proceedings of the 12th International Semantic Web Conference*. Springer. 10.1007/978-3-642-41335-3_23

- Munoz, E., Hogan, A., & Mileo, A. (2014). Using linked data to mine rdf from wikipedia tables. *Proceedings of the 7th ACM international conference on Web search and data mining (WSDM 14)*, 533–542. 10.1145/2556195.2556266
- Nagy, G., Embley, D. W., Machado, S., Seth, S., Jin, D., & Krishnamoorthy, M. (2011). Data extraction from Web tables: the devil is in the details. *International Conference on Document Analysis and Recognition (ICDAR 2011)*. 10.1109/ICDAR.2011.57
- Niiyama, C., Chung, S., Chinn, D. & Davolos, S. (2008). *Service-oriented software reengineering methodology for composite services*. TCSS 702 Design Project in Computing and Software Systems, Final, Winter.
- Perez-Castillo, R., Garcia-Rodriguez de Guzman, I., Caballero, I., Polo, M., & Piattini, M. (2009). Preciso: A reverse engineering tool to discover Web services from relational databases. *16th Working Conference on Reverse Engineering (WCRE)*, 309–310. 10.1109/WCRE.2009.21
- Scharffe, F., Atezing, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., Hamdi, F., Bihanic, L., Képéklian, G., Cotton, F., Euzenat, J., Fan, Z., Vandenbussche, P., & Vatan, B. (2012). Enabling linked data publication with the datalift platform. *Proceedings of AAAI workshop on semantic cities*.
- SCImago. (n.d.). *SCImago Journal & Country Rank*. Retrieved from <http://www.scimagojr.com>
- Takonewa, M., Mamello, T., & Paul, K. (2010). Reengineering legacy applications for SOA middleware integration. *The Southern Africa Telecommunication Networks and Applications Conference (SAT-NAC'2010)*.
- Tilley, S. R., Paul, S., & Smith, D. B. (1996). Towards a framework for program understanding. *Proceedings of the Fourth IEEE Workshop on Program Comprehension*, 19-28. 10.1109/WPC.1996.501117

ADDITIONAL READING

- Biggerstaff, T. J. (1989). Design Recovery for Maintenance and Reuse. *The Computer Journal*, 22(7), 36–49.
- Fahrner, C., & Vossen, G. (1995). A survey of Database Design Transformation Based on the Entity-Relationship Model. *Data & Knowledge Engineering*, 15(3), 213–250. doi:10.1016/0169-023X(95)00006-E
- Hainaut, J.-L. (1998). *Database Reverse Engineering*. Book, University of Namur - Institut d'Informatique.
- IEEE Std 610.12. (1990). IEEE Standard Glossary of Software Engineering Terminology. *Corrected Edition*.
- Kourouthanassis, P. E., & Giaglis, G. M. (2007). *Pervasive Information Systems. Advances in Management Information Systems*. Armonk, NY: M.E. Sharpe. ISBN: 978-0-7656-1689-0.
- Krumm, J. (2010). *Ubiquitous Computing Fundamentals*. CRC Press, Taylor & Francis Group, A CHAPMAN & HALL Book, NY. ISBN: 978-1-4200-9360-5.
- Mahapatra, R. P., & Verma, G. (2016). *Software engineering*. Khanna Book Publishing Co. (P) Ltd. ISBN: 978-93-82609-68-1.

Marciniak, J. J. (1994). *Encyclopedia of software engineering*. Wiley-Interscience New York.

Umar, A. (1997). *Application (Re)Engineering – Building Web-based Application and Dealing with Legacies*. Prentice-hall International.

KEY TERMS AND DEFINITIONS

Design Recovery: Creating design abstractions from a combination of source code, existing design documentation (if available), etc.

Forward Engineering: Moving from the high-level abstractions and logical implementation-independent designs to the low-level physical implementation of a system.

Migration: Transporting legacy systems to a newer platform (hardware or operating system) or language. For example move from a mainframe to a network-based computing environment. Migration focuses on conserving legacy systems functionalities in new environments. This needs eventually *restructuring* or *refactoring*.

Redocumentation: Creation or amendment of a semantically equivalent representation in the same relative abstraction level. It aims at recovering documentation about the subject system. It may give more explanation with alternate views such as dataflow, data structure and control flow.

Reengineering: The examination and alteration of a subject system in order to reconstitute it in a new form, and the subsequent implementation of this new form (Chikofsky¹ & Cross, *IEEE Software*, January 1990). Reengineering involves both reverse engineering and forward engineering, and it has two main objectives: (1) Software maintenance that involves *repairing* (corrective) errors or *improving* (perfective) with new required functionalities. (2) Software migration that aims at *evolving* (adaptive) systems to new environment. In particular, the reverse engineering phase may aim at *understanding* (predictive) the system.

Refactoring: Changing the source code and preserving the behavior (in the object-oriented programming).

Restructuring: The transformation from one structure to another at the same relative level of abstraction: while preserving the external behavior (functionality and semantics) of the subject system. Restructuring allows changing structure without changing functionality. It can be performed based on the knowledge of structural form; however, it does not require understanding its meaning. For example, converting a set of If statements into a Case structure, or converting *While* loop to *For* loop.

Reverse Engineering: The analysis of a subject system to identify its components and their inter-relationships, and to create its representations in another form, or at higher level of abstraction. Two subareas of reverse engineering are widely referred: *redocumentation* and *design recovery*.

Software Maintenance: Modifying a software system after delivery to correct its faults to improve its performance or other attributes, or to adapt the system to a changed environment.

ENDNOTE

¹ Elliot J. Chikofsky is described in the literature as the developer, organizer, and evangelist of software engineering automation. He is most known for his advocacy of reverse engineering and CASE technologies. He is also a consultant on management issues and information systems development methods, tools, and design techniques.

A Comprehensive Update and Performance Evaluation of Friction Factor Formulae

9

Salihu Lukman

 <https://orcid.org/0000-0002-6567-1257>

University of Hafr Al-Batin, Saudi Arabia

Isaiah Adesola Oke

 <https://orcid.org/0000-0002-7082-7682>

Obafemi Awolowo University, Nigeria

Afolabi M. Asani

Ladoke Akintola University, Nigeria

INTRODUCTION

Water and solutions are used in many technological and engineering processes (chemical, mechanical, water treatment and distribution, biomedical and pharmaceutical processes). Movement of these essential materials in the reservoir is channeled by the conservation of mass, momentum, and energy. Applications of these movements range from the low-temperature condition in gas liquefaction to high temperatures and pressures in rocket momentum systems. Actual design of all these processes requires an accurate knowledge and values of the fluid transport parameters in the pipeline and channel systems. In a specific situation, viscosity of the liquid, diameter of the pipes and tubes coupled with the friction factor are needed to calculate headloss and energy requirement in the channel and pipeline systems. A precise value of friction factor is necessary in the computation of headloss and the energy. In mechanical, chemical, water supply and biomedical processes, numerous factors are involved in the pipe network and liquefied transport systems. In medical sciences and biomedical engineering, conveyance of a physiological liquid takes place through the catheter tube into the body of a human being, which indicates that accurate computation of F_f in the catheter is a significant factor to discharge adequate liquid. Some of these significant design factors in the pipeline systems are the lengths, diameters and F_f of the pipes, water level in the reservoirs, head-discharge characteristics of the pump, water demands at different nodes and performance characteristics of different valves and minor elements in the pipe systems (Gupta and Bhawe, 2007; Özger and Yıldırım, 2009a; 2009b).

Parts of these parameters in pipe network systems remain fixed at different periods of the pipe, while some parameters would fluctuate during the life span of the pipe systems. The fluctuating parameters can be considered to be imprecise parameters. Basic equation for computing the head loss in the pipe network and pipeline system is either Hazen-Williams or Darcy-Weisbach equation that requires calculation of F_f . Darcy – Weisbach equation is expressed as follows:

$$h_L = \frac{\lambda L V^2}{2gD} \quad (1a)$$

DOI: 10.4018/978-1-7998-3479-3.ch085

$$h_L = \frac{8\lambda L Q^2}{g\pi^2 D^5} \quad (1b)$$

where; h_L is the head loss; λ is the F_f ; L is the length of the pipeline or pipe system; D is the diameter of the pipe; V is the mean velocity in the pipeline or pipe system; g is the acceleration due to gravity and Q is the flow (discharge) in the pipeline or pipe system.

Colebrook – White's formula presented an implicit expression for calculating F_f . The expressions are (Colebrook and White 1937; Colebrook, 1939; Mahendra, 2008; Mehran and Ayub, 2011; Oke *et al.*, 2015a):

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{2.51}{R_e \sqrt{\lambda}} \right) \quad (2a)$$

$$\frac{1}{\sqrt{\lambda}} = 1.74 - 2 \log_{10} \left(2 \frac{k}{D} + \frac{18.7}{R_e \sqrt{\lambda}} \right) \quad (2b)$$

$$\frac{1}{\sqrt{\lambda}} = 1.14 - 2 \log_{10} \left(1 + \frac{9.3}{R_e \frac{k}{D} \sqrt{\lambda}} \right) + 2 \log \left(\frac{D}{k} \right) \quad (2c)$$

where; λ is the friction factor; k is effective roughness size of the pipe wall and R_e is the Reynolds number.

Estimations and computations of exact values of F_f in the pipe networks are bedrock in many engineering applications (transfer of liquid and solutions through chemical reactors as well as in industrial processes which involve single-phase, double-phase). In health and medical sciences, and biomedical engineering processes, solutions flow are attained at high flow rate in blood vessels. These solutions are transported through catheter tube into the body of patients, which requires accurate value of F_f (Shaikh *et al.*, 2015). It has been shown that computation of accurate value of F_f requires an implicit equation known as Colebrook – White's formula. This implicit equation needs the use of numerical algorithms that are complex and not as fast as the explicit equations. In hydraulics of complex and supercritical pipe-flow systems it is difficult to use Colebrook – White's formula, which indicates that reliable F_{ff} are preferred. The needs for more robust and accurate F_{ff} have led researchers to propose and develop new F_{ff} . Today in mechanical, hydraulics, fluid and other engineering, there are various explicit formulae, which present a varying degree of accuracies depending on the complexity of their functional forms. It has been stated that the most complex F_{ff} usually provide higher efficiency and there is the need to establish accuracy of these F_{ff} (Özger and Yıldırım, 2009a; Shaikh *et al.*, 2015). These are F_{ff} found in literature. Moody (1944; 1947):

$$\lambda = 0.0055 \left(1 + \left[20000 \frac{k}{D} + \frac{10^6}{R_e} \right]^{\frac{1}{3}} \right). \quad (3)$$

Altshul (1952):

$$\lambda = 0.11 \left(\frac{k}{D} + \frac{68}{R_e} \right)^{0.25}. \quad (4)$$

Wood (1966):

$$\lambda = a + b R_e^{-c}, \quad (5)$$

Where; a,b and c are Wood constants as follows:

$$a = 0.094 \left(\frac{k}{D} \right)^{0.225} + 0.53 \left(\frac{k}{D} \right), \quad (6a)$$

$$b = 88.0 \left(\frac{k}{D} \right)^{0.44}, \quad (6b)$$

and

$$c = 1.62 \left(\frac{k}{D} \right)^{0.134}. \quad (6c)$$

Barr (1972) suggested an expression of the form:

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{5.1286}{(R_e)^{0.89}} \right). \quad (7)$$

Eck (1973):

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left(\left(\frac{k}{3.715D} \right) + \left(\frac{15}{R_e} \right) \right). \quad (8)$$

Churchill (1973):

$$\frac{1}{\sqrt{\lambda}} = -2 \left(\log_{10} \left(\left(\frac{k}{3.71D} \right) + \left(\frac{7}{(R_e)} \right)^{0.9} \right) \right) \quad (9)$$

Jain (1976):

$$\frac{1}{\sqrt{\lambda}} = -2 \left(\log_{10} \left(\left(\frac{k}{3.715D} \right) + \left(\frac{6.943}{R_e} \right)^{0.9} \right) \right) \quad (10)$$

Swamee and Jain (1976);

$$\lambda = \left[\frac{0.25}{\left(\log_{10} \left(\frac{k}{3.7D} + \frac{5.74}{(R_e)^{0.9}} \right) \right)^2} \right] \quad (11a)$$

$$\sqrt{\lambda} = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{5.74}{(R_e)^{0.9}} \right) \quad (11b)$$

Churchill (1977):

$$\lambda = 8 \left[\left(\frac{8}{R_e} \right)^{12} + \frac{1}{(B_i + A_i)^{\frac{3}{2}}} \right]^{\frac{1}{12}}, \quad (12)$$

$$A_i = \left(-2.457 \log_e \left(\left(\frac{k}{3.7D} \right) + \left(\frac{7}{R_e} \right)^{0.9} \right) \right)^{16}, \quad (12a)$$

$$B_i = \left(\frac{37530}{R_e} \right)^{16}. \quad (12b)$$

Chen (1979):

$$\frac{1}{\sqrt{\lambda}} = -2 \left(\log_{10} \left(\left(\frac{k}{3.7065D} \right) - \left(\frac{5.0452}{R_e} \right) \log_{10} \left(\left(\frac{1}{2.8257} \right) \left(\frac{k}{D} \right)^{1.1098} + \left(\frac{5.8506}{(R_e)^{0.8981}} \right) \right) \right) \right). \quad (13)$$

Round (1980):

$$\frac{1}{\sqrt{\lambda}} = 1.8 \log_{10} \left(\frac{R_e}{0.135 R_e \left(\frac{k}{D} \right) + 6.5} \right). \quad (14)$$

Barr (1981) suggested an expression of the form:

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{5.158 \log \left(\frac{R_e}{7} \right)}{R_e \left(1 + \frac{R_e^{0.52}}{29} \left(\frac{k}{D} \right) \right)^{0.70}} \right). \quad (15)$$

Zigrang and Sytsester(1982):

$$\lambda = \left[-2 \log_{10} \left(\frac{k}{3.77D} - \frac{5.02}{R_e} \left(\log_{10} \left[\frac{k}{3.77D} - \frac{5.02}{R_e} \left\{ \log_{10} \left(\frac{k}{3.77D} + \frac{13}{R_e} \right) \right\} \right] \right) \right) \right]^{-2} \quad (16)$$

Haaland (1983);

$$\frac{1}{\sqrt{\lambda}} = -1.8 \log_{10} \left(\frac{6.9}{R_e} + \left(\frac{k}{3.7D} \right)^{1.11} \right). \quad (17)$$

Serghides (1984),

$$\lambda = \left[A - \frac{(B_s - A_s)^2}{(C_s - 2B_s + A_s)} \right]^{-2}, \quad (18)$$

$$A_s = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{12}{R_e} \right), \quad (18a)$$

$$B_s = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{2.51A}{R_e} \right), \quad (18b)$$

$$C_s = -2 \log_{10} \left(\frac{k}{3.7D} + \frac{2.51B}{R_e} \right). \quad (18c)$$

Tsal (1989),

$$\lambda = 0.11 \left(\left(\frac{k}{D} \right) + \frac{68}{R_e} \right)^{0.25}. \quad (19)$$

If $\lambda < 0.018$ then $\lambda = 0.85\lambda + 0.0028$

Manadilli (1997):

$$\frac{1}{\sqrt{\lambda}} = -2 \left(\log_{10} \left(\left(\frac{k}{3.7D} \right) + \left(\frac{95}{(R_e)^{0.983}} \right) - \left(\frac{96.82}{R_e} \right) \right) \right). \quad (20)$$

Romeo *et al.* (2002):

$$\frac{1}{\sqrt{\lambda}} = -2 \left(\frac{\log_{10} \left(\left(\frac{k}{3.7065D} \right) - \left(\frac{5.0272}{(R_e)} \right) \right) \log_{10} \left(\frac{k}{3.827D} \right) - \left(\frac{4.567}{(R_e)} \right) \times}{\log_{10} \left(\left(\frac{k}{7.7918D} \right)^{0.9924} - \left(\left(\frac{5.3326}{(208.815 + R_e)} \right)^{0.9345} \right) \right)} \right). \quad (21)$$

$$\frac{1}{\sqrt{\lambda}} = \left(-1.8166 \log_{10} \left(\left(\frac{k}{4.1165D} \right)^{1.0593} + \left(\frac{7.0367}{(527.620 + R_e)} \right)^{0.9619} \right) \right)^{1.0131} \quad (22)$$

$$\frac{1}{\sqrt{\lambda}} = \left(-2.0064 \log_{10} \left(\left(\frac{k}{3.6795D} \right) - \left(\frac{4.8738}{R_e} \right) \log_{10} \left(\left(\frac{k}{4.2678D} \right)^{1.0287} + \left(\frac{7.4964}{(420.15 + R_e)} \right)^{0.9426} \right) \right) \right)^{0.9989} \quad (23)$$

Sonnad and Goudar (2004):

$$\frac{1}{\sqrt{\lambda}} = 0.8686 \log_e \left(\frac{0.4587 R_e}{s \left(\frac{s}{s+1} \right)} \right) \quad (24a)$$

$$s = 0.1240 R_e \left(\frac{k}{D} \right) + \log_e (0.4587 R_e). \quad (24b)$$

Sonnad and Goudar (2006):

$$\frac{1}{\sqrt{\lambda}} = 0.8686 \log_e \left(\frac{0.4587 R_e}{\left(s - 0.31 \right) \left(\frac{s}{s+1} \right)} \right). \quad (25a)$$

$$s = 0.1240 R_e \left(\frac{k}{D} \right) + \log_e (0.4587 R_e). \quad (25b)$$

Vatankhah and Kouchakzadeh (2008):

$$\frac{1}{\sqrt{\lambda}} = 0.8686 \log_e \left(\frac{0.4587 R_e}{\left(s - 0.31 \right) \left(\frac{s}{s+0.9633} \right)} \right), \quad (26a)$$

$$s = 0.1240 R_e \left(\frac{k}{D} \right) + \log_e (0.4587 R_e). \quad (26b)$$

Buzzelli (2008):

$$\frac{1}{\sqrt{\lambda}} = A_b - \frac{\left(A_b + 2 \log_{10} \left(\frac{B_b}{R_e} \right) \right)}{1 + \frac{2.18}{B_b}}, \quad (27a)$$

$$A_b = \frac{\left(\frac{0.744 \log_e (R_e) - 1.41}{\left(1 + 1.32 \sqrt{\frac{k}{D}} \right)} \right)}{\left(1 + 1.32 \sqrt{\frac{k}{D}} \right)}, \quad (27b)$$

$$B_b = \left(\frac{k}{3.7D} \right) R_e + 2.51 A_b. \quad (27c)$$

Avci and Kargoz (2009):

$$\lambda = \left(\frac{6.4}{\left(\log_e (R_e) - \log_e \left(1 + 0.001 R_e \frac{k}{D} \left(1 + 10 \sqrt{\frac{k}{D}} \right) \right) \right)^{2.4}} \right), \quad (28)$$

Evangelides *et al.* (2010):

$$\lambda = \frac{\left(\frac{0.2479 - 0.0000947 (7 - \log_{10} R_e)^4}{\left(\log_{10} \left(\frac{k}{3.615D} + \frac{7.366}{R_e^{0.9142}} \right) \right)^2} \right)}{\left(\log_{10} \left(\frac{k}{3.615D} + \frac{7.366}{R_e^{0.9142}} \right) \right)^2}. \quad (28)$$

Brkic (2011)

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left(\left(\frac{k}{3.71D} \right) + \left(\frac{2.18 S_b}{R_e} \right) \right); \quad (29a)$$

$$S_b = \log_e \left(\left(\frac{R_e}{1.816 \log_e \left(\frac{1.1 R_e}{\log_e (1 + 1.1 R_e)} \right)} \right) \right). \quad (29b)$$

Brkic (2012)

$$\frac{1}{\sqrt{\lambda}} = -2 \log_{10} \left(\left(\frac{k}{3.71 D} \right) + \left(10^{-0.4343 S_b} \right) \right); \quad (30a)$$

$$S_b = \log_e \left(\left(\frac{R_e}{1.816 \log_e \left(\frac{1.1 R_e}{\log_e (1 + 1.1 R_e)} \right)} \right) \right). \quad (30b)$$

Danish *et al.* (2011)

$$\frac{1}{2\sqrt{\lambda}} = A_d - \left(\frac{1.73718 A_d \log_e A_d}{1.73718 + A_d} \right) + \left(\frac{2.62122 A_d (\log_e A_d)^2}{(1.73718 + A_d)^3} \right) + \left(\frac{3.03568 A_d (\log_e A_d)^3}{(1.73718 + A_d)^4} \right) \quad (31a)$$

$$A_d = 4 \log_{10} (R_e) - 0.4, \quad (31b)$$

Fang *et al.* (2011) for smooth pipes:

$$\lambda = 0.25 \left[\log_{10} \left(\frac{150.39}{R_e^{0.98865}} - \frac{152.66}{R_e} \right) \right]^{-2}. \quad (32a)$$

Fang *et al.* (2011) for smooth and rough pipes:

$$\lambda = 1.613 \left[\log_e \left(0.234 \left(\frac{k}{D} \right)^{1.1007} - \frac{60.525}{R_e^{1.1105}} + \frac{56.291}{R_e^{1.0712}} \right) \right]^{-2}. \quad (32b)$$

Ghanbari *et al.* (2011) known as Ghanbari-Farshad-Rieke

$$\lambda = \left(-1.52 \log_{10} \left(\left(\frac{k}{7.21D} \right)^{1.042} + \left(\frac{2.731}{R_e} \right)^{0.9152} \right) \right)^{-2.169}. \quad (33a)$$

Shaikh *et al.* (2015)

$$\lambda = 0.25 \left[\log_{10} \left(\frac{2.51}{\alpha R_e} + \frac{k}{3.7D} \right) \right]^{-2}. \quad (34a)$$

$$\alpha = 1.14 - 2 \log_{10} \left(\frac{k}{D} \right), \quad (34b)$$

Offor and Alabi (2016a and b)

$$\lambda = \left(-2 \log_{10} \left(\left(\frac{k}{3.71D} - \frac{1.975}{R_e} \left(\log_e \left(\left(\frac{k}{3.93D} \right)^{1.092} + \left(\frac{7.627}{R_e + 395.9} \right) \right) \right) \right) \right) \right)^{-2} \quad (35)$$

Lukman and Oke (2017)

$$\lambda = \left(-1.9969 \log_{10} \left(\left(\frac{k}{3.7063D} \right) - \left(\frac{0.1800}{R_e} \right) \log_{10} \left(\left(\frac{k}{3.6636D} \right) - \left(\frac{5.3922}{R_e} \right) - \log_{10} \left(\left(\frac{k}{8.3556D} \right)^{1.7048} + \left(\frac{0.99119}{(1.000 + R_e)} \right)^{1.2744} \right) \right) \right) \right)^{-2.0017} \quad (36a)$$

$$\frac{1}{\sqrt{\lambda}} = \left(-2.0641 \log_{10} \left(\left(\frac{k}{3.4833D} \right) - \left(\frac{0.1121}{R_e} \right) \log_{10} \left(\left(\frac{k}{6.1374D} \right) - \left(\frac{1.2509}{R_e} \right) - \log_{10} \left(\left(\frac{k}{73.6856D} \right)^{0.0742} + \left(\frac{0.0121}{(1.9676 + R_e)} \right)^{10.0947} \right) \right) \right) \right)^{0.9869} \quad (36b)$$

$$\lambda = \left(\left(-2.046 \log_{10} \left(\left(\frac{k}{3.5427D} \right) - \left(\frac{0.112}{R_e} \right) \log_{10} \left(\left(\frac{k}{6.1385D} \right) - \left(\frac{1.2512}{R_e} \right) - \log_{10} \left(\left(\frac{k}{73.6864D} \right)^{0.0742} + \left(\frac{0.0121}{(1.9676 + R_e)} \right)^{10.0947} \right) \right) \right) \right)^{0.8317} \right)^{-2.3821} \quad (36c)$$

Ismail *et al.* (2017)

$$\lambda = \left(-2.2384 \log_{10} \left(\left(\frac{k}{3.7317D} \right)^{0.8919} + \left(\frac{58.2294}{R_e} \right)^{1.9458} \right) \right)^{-2} \quad (37a)$$

$$\lambda = \left(\left(-2.2118 \log_{10} \left(\left(\frac{k}{3.7562D} \right)^{0.9002} + \left(\frac{58.1100}{R_e} \right)^{1.9654} \right) \right) \right)^{1.0008} \right)^{-2} \quad (37b)$$

$$\lambda = \left(\left(-2.2639 \log_{10} \left(\left(\frac{k}{3.7621D} \right)^{0.8779} + \left(\frac{58.324}{R_e} \right)^{1.9201} \right) \right) \right)^{0.8557} \right)^{2.3416} \quad (37c)$$

Lukman *et al.* (2017a)

$$\lambda = \left(-2.1236 \log_{10} \left(\left(\frac{k}{3.6033D} \right)^{0.9531} + \left(\frac{1.0544}{(1.000 + R_e)} \right)^{1.0646} \right) \right)^{-1.9912} \quad (38a)$$

$$\frac{1}{\sqrt{\lambda}} = \left(-2.0312 \log_{10} \left(\left(\frac{k}{3.6807D} \right)^{0.9876} + \left(\frac{1.1319}{(0.9920 + R_e)} \right)^{1.1137} \right) \right)^{0.9986} \quad (38b)$$

$$\lambda = \left(\left(-2.0142 \log_{10} \left(\left(\frac{k}{3.7632D} \right)^{0.9874} + \left(\frac{1.1334}{(0.9920 + R_e)} \right)^{1.1014} \right) \right)^{0.8527} \right)^{-2.3503} \quad (38c)$$

Lukman *et al.* (2017b)

$$\lambda = \left(-1.9710 \log_{10} \left(\left(\frac{k}{3.8060D} \right) - \left(\frac{0.3013}{R_e} \right) \log_{10} \left(\left(\frac{k}{3.8025D} \right)^{5.1422} + \left(\frac{0.0899}{(61.7268 + R_e)} \right)^{0.2789} \right) \right) \right)^{-2.0123} \quad (39a)$$

$$\frac{1}{\sqrt{\lambda}} = \left(-2.0761 \log_{10} \left(\left(\frac{k}{3.4108D} \right) - \left(\frac{0.3152}{R_e} \right) \log_{10} \left(\left(\frac{k}{3.8025D} \right)^{5.1422} + \left(\frac{0.0895}{(61.7306 + R_e)} \right)^{0.2917} \right) \right) \right)^{0.9861} \quad (39b)$$

$$\lambda = \left(\left(-2.0721 \log_{10} \left(\left(\frac{k}{3.4420D} \right) - \left(\frac{0.3108}{R_e} \right) \log_{10} \left(\left(\frac{k}{3.8025D} \right)^{5.1422} + \left(\frac{0.0896}{(61.7295 + R_e)} \right)^{0.2876} \right) \right) \right)^{0.8395} \right)^{-2.3492} \quad (39c)$$

Information on adaptive neuro-fuzzy inference system and mathematically exact alternative (Özger and Yıldırım, 2009a; 2009b) and other applications can be found in Chen (1985), Jang (1993), Hong *et al.* (2002); Jin (2003); Von Bernuth and Wilson (1989); Von Bernuth (1990); Yıldırım(2009); Xiande *et al.* (2011); Winning and Coole (2013); Wikipedia (2014); Salmasi *et al.* (2012); Samadianfard (2012); Papaevangelou *et al.*(2010); Mckeen *et al.*, (2004; 2005); Čojbašića. and Brkić (2013, 2016, 2017) and Çoban (2012). Heydari *et al.* (2015) utilized surface response for F_f calculation, and its accuracy was found to be limited to a certain range. Further information on F_f can be found in Petkovic *et al.* (2013a and b), Li *et al.* (2013), Vatankhah (2014); Mustafa *et al.* (2014); Kroll *et al.*(2014); Samadianfard *et al.* (2014), Alabi and Williamson (2015), Besarati *et al.* (2015), Brikic (2014; 2016); Dejan and Carko (2016); Taler (2016), Offor and Alabi (2016a and b) and Ismail *et al.* (2018).

Essential of accurate values F_f in the selection of pipe size (water distribution systems), estimation of flows and headloss in pipes, liquid and aqueous solutions transportation analysis and in the design of potable water supply scheme are well known. Also, there are various researches and publications on calculation of accurate F_f in the pipe as well as documentations on F_{ff} for F_f computations, but update and accuracies of these F_{ff} are rare. Advancement in technology and development of the high-speed computer have made computation easy and simple, which indicates that there is the need to document F_{ff} for exact F_f calculation and provide a performance assessment of each of these F_{ff} . The principal aim of this study is to provide F_{ff} for F_f calculation and performance assessment of each of these F_{ff} with particular attention to accuracy using statistical techniques with the aim of providing error-free explicit F_{ff} .

MATERIALS AND METHOD

F_{ff} which is a function of Reynolds number (R_e) and relative roughness (R_r) were obtained from archive with the aim to update the readers and conduct performance evaluation.

Colebrook – White's formula and these F_{ff} were used to compute different F_f for $4 \times 10^3 < R_e < 1.704 \times 10^8$ and $1.0 \times 10^{-7} < R_r < 0.052$ employing Microsoft Excel Solver. Microsoft Excel solver was selected based on availability (at no additional cost) and accuracy in numerical solutions. Accuracies of these F_{ff} were evaluated using relative error; model of selection (MSC) and the Akaike Information Criterion (AIC). Procedure used in Microsoft Excel Solver can be summarized as follows (Oke *et al.*, 2014a and b, 2015a, b and c):

- Excel solver was added in the Microsoft Excel;
- Target; operation and changing cells were set based on square of difference

$$\left[\sum_{t=1}^n \lambda_t - \sum_{t=1}^n R_e \left(\frac{k}{D} \right) \right]^2 = 0 ; \text{ and}$$

- The Solver was allowed to iterate at 200 iterations with 0.005 tolerance (Figure 1).

Relative error (R_{Err}) was determined using equation (48) as follows:

$$R_{Err} = \sum_{i=1}^N \left(100 \left(\frac{Y_{expect} - Y_{cali}}{Y_{expect}} \right) \right) \quad (48)$$

Where; Y_{expect} is the F_f from Colebrook – White's formula, and Y_{cali} is the calculated F_f from other friction factor formulae.

Akaike Information Criterion: Information Criterion of Akaike (1976) allows a direct comparison of different models with a different number of parameters (Romoe *et al.*, 2002). It represents the information content of a given set of parameters by relating the coefficient of determination to the number of parameters that were required to establish the fit. The Akaike Information Criterion (AIC) was determined as follows:

$$AIC = N \left(\ln \sum_{i=1}^n (Y_{expect} - Y_{cali})^2 \right) + 2p \quad (49)$$

where; p is the total number of fixed parameters to be computed in the equation ($p = 3$); N is the total number of F_f computed ($N = 5240$).

MSC can be computed using equation (50) as follows (Babatola *et al.*, 2008; Oke, 2007):

$$MSC = \ln \frac{\sum_{i=1}^n (Y_{expect} - \bar{Y}_{expect})^2}{\sum_{i=1}^n (Y_{expect} - Y_{cali})^2} - \frac{2p}{N} \quad (50)$$

RESULTS AND DISCUSSION

Results of the study were presented in the following categories: growth and patterns of F_{ff} , accuracies of the F_{ff} .

Growth and Pattern of F_{ff} : There are 47 F_{ff} that can be used for estimation of F_f in analysis of pipe systems between 1936 and 2017. Figure 2 (a and b) presents the patterns of the growth of these F_{ff} between 1936 and 2017. The F_{ff} can be grouped into four categories of first generation F_{ff} (1930 – 1960); second generation F_{ff} (1961 to 1990); third generation F_{ff} (1991 to 2008) and recent F_{ff} (2009 to 2017). The earliest F_{ff} for pipe systems analysis related Reynolds number and relative roughness and F_f was published in 1937. The number of F_{ff} has slowly grown in the long period from 1937 to 1960. The number was 4 out of 47. The number of F_{ff} went up significantly between 1970 and 1980. It rocketed between 1980 and 2008. This indicates that F_{ff} for pipe systems has become a hot topic worldwide, because of high rate increase of related literatures and human beings urgent necessary for energy consumption reduction, effective fluid delivery and environmental pollution control. The first generation F_{ff} were Colebrook- White formulae (1937); Moody (1944), and Altshul (1952). These three formulae are implicit and complex at a certain degree difficult to use, but formed foundation and bedrock for other F_{ff} . The second generation formulae are Wood (1966); Barr (1972 and 1981); Haaland (1983); Swamee and Jain (1976); Serghide (1984), Tsal(1989), Zigrang and Sylester(1982); Churchill (1973 and 1977), Jain (1976), Chen (1979), Eck (1973), Tsal (1989) and Round (1980). These second generation's F_{ff} are explicit formulae, which are with more than two variables. These second generation formulae look little more complex than the first generation formulae, but explicit formulae which provide solutions to the implicit formula. This can be attributed to availability of computer, which aid computations and development of accurate explicit formulae numerically through iterations. The third generation F_{ff} are Vatankhah and Kouchakzadeh (2008); Buzzelli (2008); Manadilli (1997), Romeo *et al.* (2002); Sonnad and Goudar (2004 and 2006). Recent F_{ff} are Avci and Kargoz(2009); Evanleids *et al.* (2010), Brkic (2011; 2012), Danish *et al.* (2011); Fang *et al.*, (2011),Ghanbari *et al.* (2011), Shaikh *et al.*(2015), Offor and Alabi (2016a and b), Lukman and Oke (2017), Ismail *et al.* (2017), Lukman *et al.* (2017a and b). These F_{ff} (third generation and recent) are similar in form of variables and levels of parameters, but different by numerical values. These similarity and differences can be attributed to advancement in computer and technology as well as availability new numerical solutions. The number of F_{ff} grew from 1 in 1937 to 47 in 2017, which fitted into power model ($R = 0.9724$ with $R^2 = 0.9456$) and an exponential model of $R = 0.9719$ with $R^2 = 0.9446$ presented good relationship (Figure 3a). A significant correlation between the number of formulae and the year was observed with a high coefficient of determination ($R^2 = 0.9582$). Both the power and the exponential curve fitting indicated that there was a high growth rate for annual F_{ff} . The power and exponential fitting curves indicated that there are more researches on F_{ff} for fluid (gas, water and wastewater treatment etc.) transport technology were found to be $C = 0.8854X^{0.7458}$ and $C = 1.9968Exp^{0.0368X}$ respectively, Where; C is the cumulative number of F_{ff} and X is the number of years with 1936 as the base year, with coefficient of determination (CD) of 0.8589 and 0.9582 and correlation coefficient (R) of 0.9268 and 0.9789 respectively. These values of CD and R revealed that there is are good correlation for power and exponential fitting. Based on the exponential model, the number of F_{ff} in 2018 is expected to be six times more than that in 2012. The average number of F_{ff} developed per year fluctuated slightly with an overall average of 0.365 per year. Figures 3a, b and c present growth pattern of F_{ff} in various sub groups. The average number of F_{ff} developed per year were 0.129; 0.451 0.333 and 0.777 for each of the generations (sub groups). There was a high average in the

recent years, which indicated that the breadth of F_{ff} distribution as well as the broad interest in F_f research from various research angles.

The growth pattern of F_{ff} in line with various generations are as follows:

- i. Exponential relationship are $C = 1.4324Exp^{0.1024X}$, $C = 0.610Exp^{0.1306X}$, $C = 1.3859Exp^{0.3055X}$ and $C = 0.420Exp^{0.1904X}$ with CD of 0.6067, 0.9526, 0.7808 and 0.8701 for first, second, third and recent generation F_{ff} respectively; and
- ii. Powerfunctionrelationship are $C = 1.2416X^{0.5519}$, $C = 0.0277X^{2.0685}$, $C = 0.0582X^{1.6635}$ and $C = 1.1029X^{1.2751}$ with CD of 0.8567, 0.9564, 0.9582 and 0.9064 for first, second, third and recent generation F_{ff} respectively.

The grown pattern of F_{ff} in the first generation neither fits exponential nor power based on lower values of CD. The grown pattern of the second generation of F_{ff} had CD of 0.9526 and 0.9564, indicated that grown pattern fits well into power relationship than another relationship. It implies that growing and researches on F_f were of higher geometric progression at the this second stage. This can be attributed to interest in fluid transport and availability of computers. The grown patterns of the third and fourth generation formulae were similar to that of second generation formulae, which are well expressed by power fitting relationship.

Classification according to the number of the authors revealed that F_{ff} with three or more authors accounted for 37.20%, F_{ff} with two authors accounted for 22.93% and F_{ff} with single author accounted 39.87%. There was an increase in the co-authors and collaboration in the recent years (12.90% to 35.69% for three or more authors and decreased from 3.22% to 2.40% for each of single and two authors), which indicated an awareness in F_{ff} as well as the broad interest in F_f researches from various angles. Grouping of F_{ff} based on the nature of the equation revealed that the equations are of two types: simple and multiple equations. Multiple equations accounted for 21.23% and simple equation accounted for 78.77%. These ratios indicated that the developed and F_{ff} in use gave rooms for making a choice of F_{ff} to be used. From these F_{ff} , the parameters were different from each other F_{ff} , but there are general for all these F_{ff} .

Statistical evaluations of the F_{ff} : Table 1 presents values of AIC, MSC, and relative error. Based on the statistical evaluations (AIC, MSC, and relative error) of the F_{ff} , the formulae can be grouped into five categories as follows:

- a. **Perfect F_{ff} .** These are the F_{ff} with a relative error less than 1.0%, with MSC greater than 6.64 and AIC less than -34327.16. The formulae are Barr (1972), Prandtl and Nikurdse, Swamee and Jain, Serghide, Zigrang and Sylvester, Romeo et al. (2002), Brkic (2011), Eck (1973), Fang *et al* (2011, Smooth and Rough), Shaikh *et al* (2015), Offor and Alabi (2016), Lukman *et al* (2017 a and b), Ismail *et al* (2017), Lukman and Oke (2017);

Highly accurate F_{ff} . These are the F_{ff} with a relative error greater than 1.0% but less than 3.0%. The formulae are Haaland ; Wood

- b. **Moderately accurate F_{ff} .** These are the F_f with a relative error greater than 5.0%, but less than 10.0%. The formula are Moody, Churchill (1973 and 1977), Jain, Chen (1979), Buzzelli (2008), Sonnad and Goudar (2004 and 2006), Swamee and Jain (b), Barr (1981), Ghanbar et al (2011), Manadilli, Evangleids et al (2010), Vatankhah and Kouchakzadeh, Romeo et al ;

Table 1. Results of Statistical Evaluations

Friction Factor Formulae	Moody	Wood	Barr 1972	Prandtl and Nikurdse	Swamee and Jain (a)	Serghide	Tsal	Zigrang and Sylvester	Haaland
MSC	1.71	3.82	4.74	7.73	4.72	4.75	0.43	6.64	4.65
AIC	-8496.29	-19518.81	-24368.19	-40033.29	-24222.45	-24398.75	-1756.64	-34327.16	-23895.05
Relative Error	8.56	3.48	0.73	0.12	0.73	0.70	17.34	0.84	1.53
	Brkic	Eck 1973	Shakh et al 2011	Danish et al. 2011	Fang et al 2011 (Smooth)	Fang et al 2011 (Smooth and rough)	Churchill (1973)	Jain 1976	Chen 1979
MSC	4.42	4.62	10.61	-2.62	-2.39	4.74	1.98	1.99	2.04
AIC	-22690.33	-23731.87	-55122.00	14196.35	12989.76	-24375.06	-9898.59	-9917.59	-10225.71
Relative Error	0.93	0.86	0.03	98.97	84.75	0.75	8.31	8.32	8.01
	Round	Avci-Kargoz	Buzzelli	Sonnad and Goudar 2006	Manadilli	Evangleids et al	Vatankhah and Kouchakzadeh	Romeo et al 2002	Churchil 1977
MSC	1.74	-1.13	2.05	2.05	2.00	2.05	2.05	2.05	1.99
AIC	-8649.98	6426.16	-10234.37	-10258.56	-9970.90	-10235.29	-10234.34	-10230.51	-9928.67
Relative Error	11.00	45.32	7.96	7.93	8.18	7.90	7.95	7.96	8.23
	Sonnad and Goudar 2004	Swamee and Jain b	Barr 1981	Ghanbar et al 2011	Ismail et al. 2017 a	Ismail et al. 2017 b	Ismail et al. 2017 c	Romeo et al 2002 Model 1	Romeo et al 2002 Model 2
MSC	2.03	1.98	1.98	2.00	11.17	10.94	11.12	4.71	5.06
AIC	-10175.22	-9903.90	-9883.79	-10010.27	-58023.43	-56856.82	-57774.90	-24195.41	-26014.84
Relative Error	8.01	8.24	8.27	8.90	0.06	0.09	0.07	1.02	0.66
	Romeo et al 2002 Model 3	Lukman et al. 2017a a	Lukman et al. 2017b a	Lukman - Oke 2017 a	Lukman et al. 2017a b	Lukman et al. 2017b b	Lukman - Oke 2017 b	Lukman et al. 2017b c	Lukman et al. 2017a c
MSC	7.71	14.17	14.31	9.30	14.00	12.50	12.56	13.56	10.05
AIC	-39911.57	-73788.78	-74514.16	-48268.02	-72877.39	-65023.01	-65344.89	-70547.26	-52182.41
Relative Error	0.18	0.02	0.02	0.07	0.03	0.05	0.05	0.03	0.16
	Lukman - Oke 2017 c	Offor and Alabi 2016							
MSC	13.07	6.29							
AIC	-67981.93	-32497.47							
Relative Error	0.04	0.41							

c. **Low accurate F_{ff}** These are the F_f with a relative error greater than 10.0%, but less than 20.0%.

The formulae are Round (5.53%) and Tsal ;

- d. **Least accurate F_{ff} .** These are the F_f with a relative error greater than 20.0%. The formulae are Avci and Kargoz (2009), Danish *et al* (2011), and Fang *et al* (2011, smooth only).

CONCLUSION

Based on the statistical assessment of all F_{ff} , it can be concluded that:

- i. New generation F_{ff} are the best F_{ff} tools for estimating F_f in the pipeline problems based on the values of MSC, AIC, and relative error; and
- ii. there is the need to perform economics evaluation on these F_{ff} to ascertain their reliability.

REFERENCES

- Akaike, H. (1976). An information criterion (AIC). *Math. Sci*, 14, 1–5.
- Alabi, S. B., & Williamson, C. J. (2015). Neural Network-Based Model for Joint Prediction of the Newtonian and Non-Newtonian Viscosities of Black Liquor. *International Journal of Chemical Engineering and Applications*, 6(3), 195–200. doi:10.7763/IJCEA.2015.V6.480
- Altshul.A.D.(1952). Obobshchonnaya Zavisi - most dlyagidravlicheskovo raschota turboprovodor [Generalized formula for hydraulic analysis of pipes]. *Gidrotekhnicheskoye stroitelstve*, 6, 44-47.
- Avci, A., & Karagoz, I. (2009). A novel explicit equation for friction factor in smooth and rough pipes. *Journal of Fluids Engineering*, 131(6), 1–4. doi:10.1115/1.3129132
- Babatola, J. O., Oguntuase, A. M., Oke, I. A., & Ogedengbe, M. O. (2008). An Evaluation of Frictional Factors in Pipe Network Analysis Using Statistical Methods. *Environmental Engineering Science*, 25(4), 539–548. doi:10.1089/ees.2007.0086
- Barr, D. I. H. (1972). New forms of equations for the correlation of pipe resistance data. *Proc Inst Civil Eng. Tech Note*, 72(53), 383–390.
- Barr, D. I. H. (1981). Solutions of the Colebrook-White function for resistance to uniform turbulent flow. *Proceedings - Institution of Civil Engineers*, 2(71), 529–537. doi:10.1680/iicep.1981.1895
- Besarati, S. M., Myers, P. D., Covey, D. C., & Jamali, A. (2015). Modeling friction factor in pipeline flow using a GMDH-type neural network. *Cogent Engineering*, 2(1), 105–129. doi:10.1080/23311916.2015.1056929
- Brkić, D. (2011). Review of explicit approximations to the Colebrook relation for flow friction. *Journal of Petroleum Science Engineering*, 77(1), 34–48. doi:10.1016/j.petrol.2011.02.006
- Brkić, D. (2012). Can pipes be actually really that smooth? *International Journal of Refrigeration*, 35(2), 209–215. doi:10.1016/j.ijrefrig.2011.09.012

- Brkić, D. (2014). Discussion of “Gene expression programming analysis of implicit Colebrook–White equation in turbulent flow friction factor calculation” by Saeed Samadianfard. *Journal of Petroleum Science Engineering*, 124, 399–401. doi:10.1016/j.petrol.2014.06.007
- Brkić, D. (2016). A note on explicit approximations to Colebrook’s friction factor in rough pipes under highly turbulent cases. *International Journal of Heat and Mass Transfer*, 93, 513–515. doi:10.1016/j.ijheatmasstransfer.2015.08.109
- Brkić, D., & Čojbašić, Ž. (2016). Intelligent Flow Friction Estimation. *Computational Intelligence and Neuroscience*, 2016, 5242596. Advance online publication. doi:10.1155/2016/5242596
- Brkić, D., & Čojbašić, Ž. (2017). Evolutionary Optimization of Colebrook’s Turbulent Flow Friction Approximations. *Fluids*, 2(2), 15. doi:10.3390/fluids2020015
- Buzzelli, D. (2008). Calculating friction in one step. *Machine Design*, 80(12), 54–55.
- Chen, J. J. J. (1985). Systematic explicit solutions of Prandtl and Colebrook–White equations for pipe flow. *Proc Inst Civil Eng 2. Tech Note*, 431(79), 383–389.
- Chen, N. H. (1979). An explicit equation for friction factor in pipe. *Industrial & Engineering Chemistry Fundamentals*, 18(3), 296–304. doi:10.1021/i160071a019
- Churchill, S. W. (1973). Empirical expressions for the shear stressing turbulent flow in commercial pipe. *AIChE Journal. American Institute of Chemical Engineers*, 19(2), 375–376. doi:10.1002/aic.690190228
- Churchill, S. W. (1977). Friction factor equation spans all fluid-flow regimes. *Chemical Engineering (Albany, N.Y.)*, 84(24), 91–92.
- Çoban, M. T. (2012). Error analysis of non-iterative Friction factor formulas relative to Colebrook-White equation For the calculation of pressure Drop in pipes. *Journal of Naval Science and Engineering.*, 8(1), 1–13.
- Čojbašić, Ž., & Brkić, D. (2013). Very accurate explicit approximations for calculation of the Colebrook friction factor. *International Journal of Mechanical Sciences*, 67, 10–13. doi:10.1016/j.ijmecsci.2012.11.017
- Colebrook, C. F. (1939). Turbulent flow in pipes, with particular reference to the transition region between the smooth and rough pipe laws. *Journal of the Institution of Civil Engineers*, 11(4), 133–156. doi:10.1680/ijoti.1939.13150
- Colebrook, C. F., & White, C. M. (1937). Experiments with fluid friction in roughened pipes. *Proc R Soc A Mat*, 161(904), 367–381.
- Danish, M., Kumar, S., & Kumar, S. (2011). Approximate explicit analytical expressions of friction factor for flow of Bingham fluids in smooth pipes using Adomian decomposition method. *Communications in Nonlinear Science and Numerical Simulation*, 16(1), 239–251. doi:10.1016/j.cnsns.2010.03.013
- Dejan, B. D., & Carko, S. (2016). Intelligent Flow Friction Estimation. *Hindawi Publishing Corporation Computational Intelligence and Neuroscience ID*, 5242596. Advance online publication. doi:10.1155/2016/5242596
- Eck, B. (1973). *Technische Stromungslehre* (1st ed.). Springer.

Evangelides, C., Papaevangelou, G., & Tzimopoulos, C. (2010). *A New Explicit Relation for the Friction Factor Coefficient in the Darcy–Weisbach Equation, Protection and Restoration of the Environment*. Corfu, Greece: University of Ioannina Greece and Stevens Institute of Technology.

Fang, X., Xu, Y., & Zhanru, Z. (2011). New correlations of single-phase friction factor for turbulent pipe flow and evaluation of existing single-phase friction factor correlations. *Nuclear Engineering and Design*, 241(3), 897–902.

Ghanbari, A., Farshad, F. F., & Rieke, H. H. (2011). Newly developed friction factor correlation for pipe flow and flow assurance. *Journal of Chemical Engineering and Materials Science*, 2(6), 83–86.

Gupta, R., & Bhawe, P. R. (2007). Fuzzy parameters in pipe network analysis. *Civil Engineering and Environmental Systems*, 24(1), 33–54. doi:10.1080/10286600601024822

Haaland, S. E. (1983). Simple and explicit formulas for the friction factor in turbulent pipe flow. *Journal of Fluids Engineering*, 105–107.

Heydari, A., Elhameh, N., & Fatemeh, P. (2015). Explicit Determinations of the Colebrook Equation for the Flow Friction Factor by Statistical Analysis. *Chemical Engineering & Technology*, 38(8), 1387–1396. doi:10.1002/ceat.201400590

Hong, Y. S., Rosen, M. R., & Reeves, R. R. (2002). Dynamic fuzzy modeling of storm water infiltration in urban fractured aquifers. *Journal of Hydrologic Engineering*, 7(5), 380–391. doi:10.1061/(ASCE)1084-0699(2002)7:5(380)

Ismail, A., Lukman, S., & Oke, I. A (2017) A Numerical Solution Of Colebrook White Friction Factor Formula. In *Pipe Network Environmental Design and Management International Conference (EDMIC 2017)*. Faculty of Environmental Design and Management, Obafemi Awolowo University.

Ismail, A., & Lukman, S. And Oke, I. A. (2018). Statistical Evaluation Of Effects Of Reynolds Number And Relative Roughness On Friction Factor In Pipes Under Turbulent Flow Using Factorial Experiment. *Nigerian Journal of Engineering*, 24(2), 75–91.

Jain, A. K. (1976). Accurate explicit equations for friction factor. *J Hydraulic Div, ASCE*, 102(5), 674–677.

Jang, J.-S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3), 665–685. doi:10.1109/21.256541

Jin, Y. (2003). *Advanced fuzzy systems design and applications*. Physica. doi:10.1007/978-3-7908-1771-3

Kroll, A., & Schulte, H. (2014). Benchmark problems for nonlinear system identification and control using soft computing methods: Need and overview. *Applied Soft Computing*, 25, 496–513. doi:10.1016/j.asoc.2014.08.034

Li, L. K., Shao, S., & Yiu, K.-F. C. (2013). A new optimization algorithm for single hidden layer feed-forward neural networks. *Applied Soft Computing*, 13(5), 2857–2862. doi:10.1016/j.asoc.2012.04.034

Lukman, S., Adeosun, O., Bello, P. U., & Oke, I. A. (2017a). An Improved And Simple Fiction Factor Formulae Analysis. *Proceedings Of The 9th Annual Conference Of The Institute Of Ecology And Environmental Studies*, 138–147.

Lukman, S., Adeosun, O. O., Bello, P. U., & Oke, I. A. (2017b). Rainwater Improvement And Harvest: Error Free Friction Factor Formulae For Calculating F_f In The Pipe Flow Systems. *Proceedings of the 8th Nigerian Association of Hydrological Sciences (NAHS, ABUJA, 2017)*, 84-92.

Lukman, S., & Oke, I. A. (2017). Accurate Solutions Of Colebrook- White's Friction Factor Formulae. *Nigerian Journal of Technology*, 36(4), 1039–1048. doi:10.4314/njt.v36i4.8

Mahendra, P. V. (2008). Moody Chart: An Activex Component To Calculate Frictional Factor For Fluid Flow In Pipelines. *Proceedings, Thirty-Third Workshop on Geothermal Reservoir Engineering Stanford University*.

Manadilli, G. (1997). Replace implicit equations with signomial functions. *Chemical Engineering (Albany, N.Y.)*, 104(8), 129–135.

Mckee, B. J., Swanson, C. J., Zagarola, M. V., Donnelly, R. J., & Smits, A. J. (2004). Friction factors for smooth pipe flow. *Journal of Fluid Mechanics*, 511, 41–44. doi:10.1017/S0022112004009796

Mckee, B. J., Zagarola, M. V., & Smits, A. J. (2005). A new friction factor relationship for fully developed pipe flow. *Journal of Fluid Mechanics*, 538(-1), 429–443. doi:10.1017/S0022112005005501

Mehran, Y., & Ayub, B. (2011). Estimation of Friction Factor in Pipe Flow Using Artificial Neural Networks. *Canadian Journal on Automation. Control and Intelligent Systems*, 2(4), 52–56.

Moody, L. F. (1944). Friction factors for pipe flow. *Trans ASME*, 66, 671–684.

Moody, L. F. (1947). An approximate formula for pipe friction factors. *Trans ASME*, 69, 1005–1006.

Mustafa, A., Oguz, E. T., & Mustafa, T. C. (2014). A review of non iterative friction factor correlations for the calculation of pressure drop in pipes. *Bitlis Eren Univ J Sci and Technol*, 4(1), 1–8.

Offor, U. H., & Alabi, S. B. (2016a). Performance Evaluation Of The Explicit Approximations Of The Implicit Colebrook Equation. *IJRET. International Journal of Research in Engineering and Technology*, 05(08), 1–12. doi:10.15623/ijret.2016.0508001

Offor, U. H., & Alabi, S. B. (2016b). An Accurate and Computationally Efficient Friction Factor Model. *Advances in Chemical Engineering Science*, 6(03), 237–245. doi:10.4236/aces.2016.63024

Oke, I. A. (2007). Reliability and Statistical Assessment of Methods For Pipe Network Analysis. *Environmental Engineering Science*, 24(10), 1481–1489. doi:10.1089/ees.2007.0001

Oke, I. A., Adie, D. B., Ismail, A., Lukman, S., Igboro, S. B., & Sanni, M. I. (2014a). Computer Simulation In The Development And Optimization Of Carbon Resin Electrodes For Water And Wastewater Treatment Electrochemically. *Ife Journal of Science*, 16(2), 227–239.

Oke, I. A., Ismail, A., Lukman, S., Adie, D. B., Adeosun, O. O., Umaru, A. B., & Nwude, M. O. (2015b). Statistical Evaluation Of Mathematical Methods In Solving Linear Theory Problems: Design Of Water Distribution Systems. *Ife Journal of Science*, 17(2), 255–267.

Oke, I. A., Ismail, A., Ojo, S. O., Adeosun, O. O., Nwude, M. O., & Olatunji, S. A. (2015c). Enhanced Solutions in Pipe Network Analysis through Utilization of Microsoft Excel Solver. *Journal of Env. Design Management.*, 7(1), 15–28.

Oke, I. A., & Obijole, O. A. (2014b). *Assessment of Techniques For Solving Factorial Experiment: Adsorption Factors*. A paper presented at 7th Annual Conference on the Environmental jointly organized by The Institute of Ecology and Environmental Studies, Obafemi Awolowo University.

Oke, I. A., Ojo, S. O., & Adeosun, O. O. (2015a). Performance Evaluation For friction Factor Formulae Using Colebrook-White As Reference. *Ife Journal of Science*, 17(1), 75–86.

Özger, M., & Yıldırım, G. (2009a). Neuro-fuzzy approach in estimating Hazen–Williams friction coefficient for small-diameter polyethylene pipes. *Advances in Engineering Software*, 40(8), 593–599. doi:10.1016/j.advengsoft.2008.11.001

Özger, M., & Yıldırım, G. (2009b). Determining turbulent flow friction coefficient using adaptive neuro-fuzzy computing technique. *Advances in Engineering Software*, 40(4), 281–287. doi:10.1016/j.advengsoft.2008.04.006

Papaevangelou, G., Evangelides, C., & Tzimopoulos, C. (2010). A new explicit equation for the friction coefficient in the Darcy-Weisbach equation. *Proceedings of the Tenth Conference on Protection and Restoration of the Environment: PRE10*, 1-7.

Petkovi'c, D., Čojbašić, Ž., & Lukić, S. (2013). Adaptive neuro fuzzy selection of heart rate variability parameters affected by autonomic nervous system. *Expert Systems with Applications*, 40(11), 4490–4495. doi:10.1016/j.eswa.2013.01.055

Petkovi'c, D., Pavlović, N. D., Čojbašić, Ž., & Pavlović, N. T. (2013). Adaptive neuro fuzzy estimation of underactuated robotic gripper contact forces. *Expert Systems with Applications*, 40(1), 281–286. doi:10.1016/j.eswa.2012.07.076

Romeo, E., Royo, C., & Monzon, A. (2002). Improved explicit equations for estimation of the friction factor in rough and smooth pipes. *Chemical Engineering Journal*, 86(3), 369–374. doi:10.1016/S1385-8947(01)00254-6

Round, G. F. (1980). An explicit approximation for the friction-factor Reynolds number relation for rough and smooth pipes. *Canadian Journal of Chemical Engineering*, 58(1), 122–126. doi:10.1002/cjce.5450580119

Salmasi, F., Khatibi, R., & Ghorbani, A. M. (2012). A study of friction factor formulation in pipes using artificial intelligence techniques and explicit equations. *Turkish J. Eng. Env. Sci*, 36, 121–138.

Samadianfard, S. (2012). Gene expression programming analysis of implicit Colebrook-White equation in turbulent flow friction factor calculation. *Journal of Petroleum Science Engineering*, 92-93, 48–55. doi:10.1016/j.petrol.2012.06.005

Samadianfard, S., Taghi Sattari, M., Kisi, O., & Kazemi, H. (2014). Determining flow friction factor in irrigation pipes using data mining and artificial intelligence approaches. *Applied Artificial Intelligence*, 28(8), 793–813. doi:10.1080/08839514.2014.952923

Serghides, T. K. (1984). Estimate friction factor accurately. *Chemical Engineering (Albany, N.Y.)*, 91, 63–64.

Shaikh, M. M., Shafiq-ur-Rehman, M., & Asim, I. W. (2015). A new explicit approximation to Colebrook's friction factor in rough pipes under highly turbulent cases. *International Journal of Heat and Mass Transfer*, 88, 538–543. doi:10.1016/j.ijheatmasstransfer.2015.05.006

Sonnad, J. R., & Goudar, C. T. (2004). Constraints for using Lambert W function-based explicit Colebrook–White equation. *Journal of Hydraulic Engineering*, 130(9), 929–931. doi:10.1061/(ASCE)0733-9429(2004)130:9(929)

Sonnad, J. R., & Goudar, C. T. (2006). Turbulent flow friction factor calculation using a mathematically exact alternative to the Colebrook-White equation. *Journal of Hydraulic Engineering*, 132(8), 863–867. doi:10.1061/(ASCE)0733-9429(2006)132:8(863)

Swamee, P. K., & Jain, A. K. (1976). Explicit equations for pipe-flow problems. *Journal of the Hydraulics Division*, 102(HY5), 657–664.

Taler, D. (2016). Determining velocity and friction factor for turbulent flow in smooth tubes. *International Journal of Thermal Sciences*, 105, 109–122. doi:10.1016/j.ijthermalsci.2016.02.011

Timung, S., & Mandal, T. K. (2013). Prediction of flow pattern of gas–liquid flow through circular microchannel using probabilistic neural network. *Applied Soft Computing*, 13(4), 1674–1685. doi:10.1016/j.asoc.2013.01.011

Torbjörn, L., Seifert, E., Abramo, L., Bernt, T., Nystroöm, A., Pettersen, J., & Bergman, R. (1998). Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2), 3–40. doi:10.1016/S0169-7439(98)00065-3

Tsal, R. J. (1989). Altshul-Tsal friction factor equation, Heating. *Piping and Air Conditioning*. ASCE, 102, 657–664.

Vatankhah, A. R. (2014). Comment on “Gene expression programming analysis of implicit Colebrook–White equation in turbulent flow friction factor calculation. *Journal of Petroleum Science Engineering*, 124, 402–405. doi:10.1016/j.petrol.2013.12.001

Vatankhah, A. R., & Kouchakzadeh, S. (2008). Discussion of “Turbulent flow friction factor calculation using a mathematically exact alternative to the Colebrook-White equation” by Jagadeesh R. Sonnad and Chetan T. Goudar. *Journal of Hydraulic Engineering*, 134(8), 1187–1194. doi:10.1061/(ASCE)0733-9429(2008)134:8(1187)

Von Bernuth, R. D. (1990). Simple and accurate friction loss equation for plastic pipe. *Journal of Irrigation and Drainage Engineering*, 116(2), 294–298. doi:10.1061/(ASCE)0733-9437(1990)116:2(294)

Von Bernuth, R. D., & Wilson, T. (1989). Friction factors for small diameter plastic pipes. *Journal of Hydraulic Engineering*, 115(2), 183–192. doi:10.1061/(ASCE)0733-9429(1989)115:2(183)

Wikipedia. (2014). *Friction factor formulae*. http://en.wikipedia.org/wiki/Darcy_friction_factor_formulae

Winning, H. K., & Coole, T. (2013). Explicit Friction Factor Accuracy and Computational Efficiency for Turbulent Flow in Pipes Flow. *Turbulence and Combustion*, 90(1), 1–27. doi:10.1007/10494-012-9419-7

Wood, D. J. (1966). An explicit friction factor relationship. *Civil Eng, ASCE*, 1966, 60–65.

Xiande, F., Yu, X., & Zhanru, Z. (2011). New correlations of single-phase friction factor for turbulent pipe flow and evaluation of existing single-phase friction factor correlations. *Nuclear Engineering and Design*, 241(3), 897–902. doi:10.1016/j.nucengdes.2010.12.019

Yıldırım, G. (2009). Computer-based analysis of explicit approximations to the implicit Colebrook-White equation in turbulent flow friction factor calculation. *Advances in Engineering Software*, 40(11), 1183–1190. doi:10.1016/j.advengsoft.2009.04.004

Zigrang, D. J., & Sylvester, N. D. (1982). Explicit approximations to the solution of Colebrook's friction factor equation. *AIChE Journal. American Institute of Chemical Engineers*, 28(3), 514–515. doi:10.1002/aic.690280323

The Role of Two-Dimensional Materials in Electromagnetic Interference Shielding

Rafael Vargas-Bernal

 <https://orcid.org/0000-0003-4865-4575>

Instituto Tecnológico Superior de Irapuato, Mexico

INTRODUCTION

The extensive use of electronic and wireless technologies has led to the need to develop materials that present high performance shielding against electromagnetic interference (EMI), which avoid pollution by electromagnetic waves that affect the operation and life time of electronic devices as well as the Human health (Song, 2017; Wang, 2017; Liu, 2019). Commonly, metallic materials are used in practical ways to increase the shielding effectiveness (SE) through an appropriately designed assembly process. Unfortunately, the high density of devices that require it and the poor environmental stability of metals have impeded their massive use. In addition, materials with light weight and good chemical stability are also required for applications in the automotive, aerospace and electronics industries.

Research in the area of materials engineering has introduced two-dimensional materials in the last decades as an alternative option to increase the efficiency of electromagnetic shielding. Two-dimensional materials are defined as crystalline materials of a single layer of two-dimensionally arranged allotropic atoms based on elements or compounds (two or more covalently bonded elements). In the following paragraphs, the fundamental role that these materials play in the future of electromagnetic shielding is introduced. The advent of two-dimensional materials in the area of electromagnetic interference will mitigate the harmful effects of excessive use of electronic and telecommunications devices through electromagnetic shielding for commercially, industrially and militarily used systems. In this way, the manufacture of electronic systems must include an electronic packaging that involves the application of two-dimensional materials used as electronic fillers in composite materials based on polymeric, ceramic or metallic matrices. From a commercial point of view, the integration of two-dimensional materials will provide a high added value to products that include more efficient electromagnetic shielding than that achieved with current technologies.

Carbon-based nanomaterials such as graphene and carbon nanotubes have recently been introduced to control electromagnetic interference or mitigate contamination by electromagnetic radiation, and because they offer characteristics such as low density, high electrical conductivity and good chemical resistance (Dhakate, 2015; Xu, 2018; Zhao, 2018a). Of these nanomaterials, graphene is a two-dimensional carbon material with a higher electrical conductivity than carbon nanotubes, which further increases the shielding effectiveness (Bhattacharjee, 2017; Liu, 2019). To use the multifunctional properties of these nanomaterials in practical applications, it is necessary to implement nanocomposites that contain them regularly embedded in polymer matrices. Two-dimensional materials must be incorporated in small quantities and form efficient conductive networks within the composite material that will be formed by exploiting their high intrinsic electrical conductivity, large aspect ratio, good dispersion, and homogeneous distribution.

Carbonitrides and/or carbides based on transition metals called MXenes represent other highly conductive two-dimensional nanomaterials that can be used to provide electromagnetic interference

DOI: 10.4018/978-1-7998-3479-3.ch086

shielding when incorporated into polymeric materials such as polystyrene (Sun, 2017a; Cao, 2018; Fan, 2018; Liu, 2018; Raagulan, 2018; Cao, 2019). Additionally, two-dimensional materials are attractive for their light weight, flexibility, hydrophobicity, good durability, strong absorption capacity, and shielding against electromagnetic interference up to 70 dB that are not offered by bulk materials. These two-dimensional materials must be used functionalized or used in combination with metals (Liu, 2019) or ceramics (Sharif, 2017; Jin, 2018; Ru, 2018) embedded in polymer matrices based on thermoplastics, thermosets or conductive polymers. Other two-dimensional materials that exhibit shielding properties are black phosphorus (BP), transition metal dichalcogenides (MoS_2), and carbon nitride (C_3N_4). Multiple two-dimensional materials with van der Waals structure or heterostructures have also been used (Zhao, 2018b).

The purpose of this chapter is to describe the impact that two-dimensional materials (or 2D materials) are having on the development of materials used for electromagnetic interference shielding, particularly the impulse of materials such as graphene and MXenes. The advances in the last decade are analyzed and alternatives are proposed that will come in the next decades. The shielding mechanisms presented by two-dimensional materials are analyzed in detail and the specific applications in which these materials can be used are presented.

BACKGROUND

In this section, the chapter describes the basic concepts associated with the technical principles on which the shielding to electromagnetic interference is based. The overly exhaustive use of technologies operating in the high-frequency range has led to a new type of pollution called electromagnetic interference (EMI) (Bhattacharjee, 2017). Any device that transmit, distributes or uses electrical energy produces electromagnetic interference, which can have a damaging impact on the performance of other devices or the surrounding environment. This problem has been accentuated more because the electronic devices have reduced the size of their components and operate at much faster speeds, which can encourage the malfunction and/or degradation of the devices. This electromagnetic pollution has not been completely restricted, since it is also affecting the human being as well as the surrounding environment due to the absence of adequate shielding. The largest problems of electromagnetic interference are observed in applications where large amounts of data are used such as electronic equipment, industrial equipment and communication systems (Cao, 2019). Systems present two types of interference: electronics and radiation sources. Electromagnetic shielding involves the blocking of electromagnetic radiation so that it does not pass through a blocking or shielding material (Dhatake, 2015). Electromagnetic interference shielding is used in applications such as the radar absorption, aviation, portable electronic devices, camouflage materials, automotive systems, aerospace systems, and military systems (Raagulan, 2018). Materials that are capable of absorbing and attenuating incident microwaves used in military applications are called microwave absorbing materials (Wang, 2018). Since the late 1940s, many researchers around the world have turned their attention to studying microwave absorbing materials with two specific properties: lightweight and high absorption efficiency. In addition, these materials must have matching properties of optimal impedance (to ensure that the electromagnetic waves enter the material) and high attenuation constant (to consume the energy of the electromagnetic wave). An excellent electromagnetic wave absorbing material must be capable of absorbing waves over a wide range of frequencies, having a low density, lightweight, large surface area, and good thermal and chemical stability (Prasad, 2018).

An option to respond to electromagnetic interference (EMI) is to design and manufacture materials that can reduce or restrict the undesired effect of these electromagnetic waves (Bhattacharjee, 2017). The desired properties for these materials include a low reflection loss, a broad attenuation band, lightweight and economic viability. A successful electromagnetic shielding material must both decrease unwanted emissions and defend the device or human being from these signals. The primary function of electromagnetic shielding is to reflect the electromagnetic radiation using the charge carriers of the materials that interact with electromagnetic fields, so it is expected that the material used in the shield has a semiconductor or metallic electrical behavior. As a second function, it is required that the electrical and magnetic dipoles of the material absorb the electromagnetic radiation during the interaction with the electromagnetic fields. And, as a third function of the shielding material, within its structure there is the possibility that multiple internal reflections are generated, which scatter or absorb the electromagnetic waves due to the presence of dispersion centers or defects. The material that commonly perform this function are metals, but in the case of smaller devices their inclusion increases their weight and make them susceptible to chemical corrosion. In addition, metals are difficult and expensive to process and have low specific shielding capabilities (Dhakate, 2015). As alternative materials, polymer matrix composites that make use of conductive fillers and/or materials with magnetic properties have been proposed to replace metals, which have performed similarly but not as effectively until the advent of the two-dimensional materials. The shielding effectiveness of the polymer composite depends on the intrinsic electrical conductivity, dielectric constant, magnetic permeability, aspect ratio, and content of the conductive filler. The main disadvantage of nanocomposites is its tendency to reflect electromagnetic radiation instead of absorbing it. Another attractive option has been the use of conductive polymers that have finite electrical conductivity and no transparency to microwaves.

The shielding efficiency or shielding effectiveness (SE) of a material determines its ability to attenuate the incident electromagnetic radiation and is expressed in decibels (dB) (Bhattacharjee, 2017). The greater the shielding effectiveness, the smaller the amount of energy that passes through it. The shielding is most effective when most of the electromagnetic radiation is either reflected or absorbed by the shield. In the case of reflection, its performance is governed by the surface reflection of the free charge carriers found on the surface of the material, as well as by the multiple internal reflections produced by the adjacent electrically conductive layers. On the other hand, the absorption of the electromagnetic radiation in the shielding material is governed by different loss mechanisms related to the electric and magnetic polarization processes within the material. Therefore, the total shielding effectiveness (SE_T) can be expressed mathematically as the sum of the shielding effectiveness to reflection (SE_R), to multiple reflections (SE_{MR}) and to absorption (SE_A):

$$SE_T = SE_R + SE_{MR} + SE_A = \log_{10} \left(\frac{P_i}{P_t} \right) \quad (1)$$

where P_i and P_t represents the incident and transmitted power radiation, respectively. Because the transmitted power is expected to be much smaller than the incident power, the value of the effectiveness is negative, and the higher this value is, the better the electromagnetic shielding obtained. Regularly, the losses associated with multiple internal reflections can be ignored when the effectiveness of the shielding material is greater than -10 dB (Dhakate, 2015). Effectiveness by reflection can be determined mathematically by the equation:

$$SE_R = -10 \log \left(\frac{\sigma_T}{16\omega\epsilon_0\mu_r} \right) \quad (2)$$

where σ_T is the total electrical conductivity of the shielding material, ω is the angular frequency of the electromagnetic radiation, ϵ_0 is the electrical permittivity of the vacuum and μ_r is the relative permeability of the shielding material. From Equation 2, it can be deduced that the reflection is directly proportional to the ratio of the electrical conductivity and magnetic permeability of the material and inversely proportional to the electromagnetic signal's frequency. Effectiveness by absorption can be determined mathematically by the equation:

$$SE_A = -8.68t \sqrt{\frac{\sigma_T \omega \mu_r}{2}} \quad (3)$$

where t is the thickness of the shielding material. From Equation 3, it can be deduced that the absorption is directly proportional to the thickness, electrical conductivity and magnetic permeability of the shielding material as well as to the angular frequency of the incident radiation. In addition, according to Equation 3, the thicker the shielding material is, the greater the absorption of the applied electromagnetic wave. Because the total electrical conductivity can be expressed as

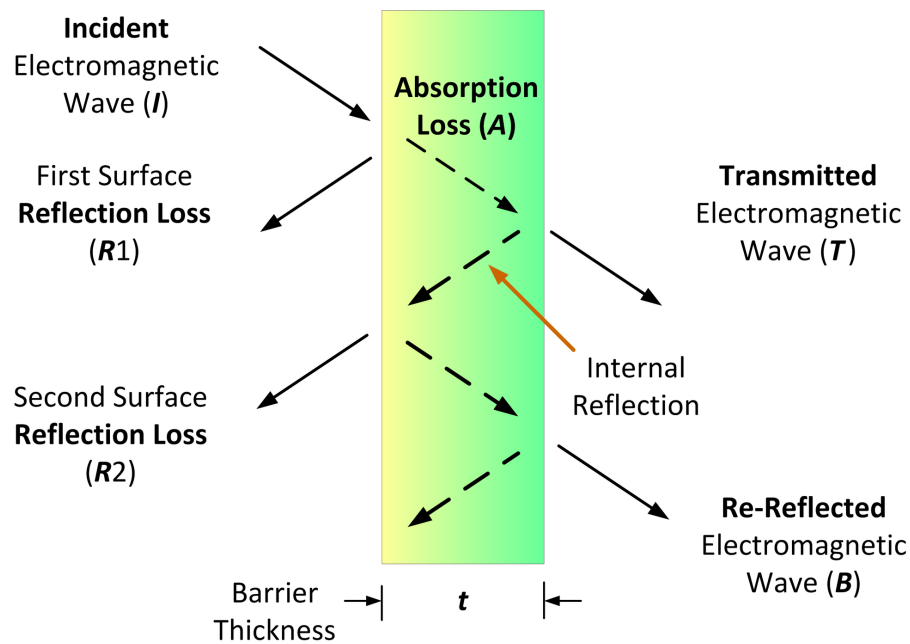
$$\sigma_T = \omega\epsilon_0\epsilon'' \quad (4)$$

where

$$\epsilon_T = \epsilon' - j\epsilon'' \quad (5)$$

Therefore, the electrical conductivity of the shielding material depends on its complex electrical permittivity. Further studies have shown that the shielding mechanism depends exclusively on the losses, which are improved by increasing the transmission routes of the incident waves that involve only the values of relative permittivity and permeability of the shielding material. In this way, the design and manufacture of materials that offer better routes for reflection, multiple reflections and absorption, which reach to dissipate the effect of electromagnetic radiation, is a priority so that these can be applied as shielding materials. Physicochemical properties such as electrical conductivity, electromagnetic polarization, crystalline structure, aspect ratio, thickness, type of material and fillers influence in the efficiency of specific electromagnetic shielding (Raagulan, 2018). Multiple internal reflections are favored by controlling thickness, skin depth, absorption of electric and magnetic dipoles, use of multi-layer materials and scattering centers. The electromagnetic reflection mechanism is favored with a high electrical conductivity. The new materials used as electromagnetic shielding materials must then be able to tune their performance by controlling their thickness and improving the mechanical properties to be able to adapt as coatings. The main mechanisms that occur in materials due to the incidence of electromagnetic waves on them are illustrated in Figure 1.

Figure 1. Mechanisms of interaction of electromagnetic waves with materials



FOCUS OF THE ARTICLE

In this section, the chapter describes the basic concepts associated with two-dimensional materials and the relevant properties that these materials possess to develop electromagnetic shielding. Two-dimensional materials have been considered as emerging materials in innumerable applications thanks to their unique electrical, optical, thermal, magnetic, and mechanical properties. The main types of materials that can be used in electromagnetic shielding are dielectric loss materials, magnetic loss materials and electromagnetic composites. The main morphologies based on nanomaterials for shielding that have been used so far are nanoparticles, nanowires, nanobars, nanotubes and nanosheets (Cao, 2019). Among the promising two-dimensional materials for electromagnetic interference shielding are materials such as graphene, MXenes, transition metal dichalcogenides, black phosphorus and graphitic carbon nitride ($g\text{-C}_3\text{N}_4$). In addition to electromagnetic interference shielding, two-dimensional materials have potential to be used in electromagnetic fields, microwave absorption, information security, electromagnetic protection, and electromagnetic wave imaging. With the introduction of two-dimensional materials, researchers around the world are working on the preferential assembly of these materials in multilayered or sandwich structure architectures. These architectures offer better electromagnetic shielding properties, lower percolation threshold, flexibility and impact toughness, which are other very important qualities during the manufacture and operation of these materials (Bhattacharjee, 2017). The electronic behavior of two-dimensional materials during electromagnetic shielding is strongly related to their crystalline structure and their electronic states. Each two-different material has a different electronic structure, energy-band structure, density of states, electronic spin, charge difference density, phonon vibration, etc. very different. The electrical conductivity can be increased by decreasing the band gap, while the lattice vibration can reduce it when producing carrier scattering. The defects and groups on the materials modify the charge density giving rise to dipoles which generate relaxation if the electromagnetic field of the interference is alternating. The magnetic spin dipole comes from the location and position

of electrons which can move due to the presence of electromagnetic radiation. The electromagnetic response of two-dimensional materials is derived from their electron transport, dipole relaxation, magnetic resonance and Eddy current (Cao, 2019). For a shielding material (common in a two-dimensional material) to absorb electromagnetic radiation, it must have a dielectric constant as close as possible to that of air (Dhakate, 2015).

MXenes are a unique family of two-dimensional materials based on transition metal carbides and/or nitrides with the chemical formula $M_{n+1}X_nT_x$, where M is a transition metal (Ti, Zr, V, Nb, Ta or Mo) and X is a chalcogen (S or O) (Bhattacharjee, 2017; Anasori, 2019). Because these materials are obtained in aqueous media during their synthesis, they can be terminated with different surface moieties (T_x), such as a mixture of OH^- , O^{2-} , and F^- . In addition, these materials have metallic conductivity, good mechanical properties, and are hydrophilic, which makes them attractive to form polymer composite materials (Shahzad, 2016). Polymer-MXene composites have an improved tensile strength and good electrical conductivity even at low percolation thresholds. The properties described above make these materials an emerging alternative to apply in shielding electromagnetic interference. MXene materials perform better on electromagnetic shielding than many other metals and carbon-based materials (Sun, 2017a). In the MXenes group, the $\text{Ti}_3\text{C}_2\text{T}_x$ has been the most explored and has an electrical conductivity of 650 000 S/m. The advantages of MXenes over other shielding materials are its laminated morphology and its unique physical and chemical properties.

Carbon-based materials have always shown the potential to absorb microwaves in almost all their morphologies (black carbon, carbon nanotubes, carbon fibers and graphene), but graphene stands out among them (Sharon, 2015; Wang, 2018). Graphene sheets are two-dimensional layers of an atom thick with carbon atoms with sp^2 bonds (Dhakate, 2015). Graphene exhibits high electrical conductivity, large surface area, excessively low density, thin thickness and fascinating advantages for developing radar absorbing materials or as shielding materials for electromagnetic interference. The advantages of composite materials based on graphene are its low cost, high aspect ratio, and lightweight. Due to its low density and resistance to corrosion, graphene and reduced graphene oxide, also two-dimensional materials, have recently been proposed as alternative materials to develop electromagnetic shielding. The presence of functional groups and defects in graphene can produce asymmetric charge density differences producing electrical dipoles and if these are subjected to alternating electromagnetic fields, the dipoles will break and generate orientation rotation, and convert the electromagnetic energy into thermal energy during the relaxation (Cao, 2019). Pure graphene cannot satisfy the impedance matching condition required for microwave absorption due to its high electrical conductivity and electromagnetic parameters (Hao, 2018). In order to use graphene in microwave absorbing materials, it is necessary to design functional composite materials with other active second phases to reduce their electrical conductivity and to achieve large-scale production. Although the interfacial interactions in these composites improve the microwave absorption properties, the second phases, usually based on metals and metal oxides, give rise to materials with high densities, which makes them impractical.

Transition metal dichalcogenides (TMDs) monolayers are an option of two-dimensional materials that exhibit a semiconductor behavior with chemical formula MX_2 , where M is a transition metal atom (Mo, W, etc.) and X a chalcogen atom (S, Se, or Te). These materials are structured so that a layer of transition metal atoms is sandwiched between two layers of chalcogen atoms (Kolobov, 2016; Wang, 2018). Among the electronic properties that dichalcogenides present are a direct band gap, its structure has no inversion center which increases the degree of freedom of charge carriers, and a strong spin-orbit coupling leads to a splitting in hundreds of meV the valence band and in a few meV to the conduction band, which allows to control the electron spin minutely. Within its large family, molybdenum disulfide

(MoS₂) exhibits high electron mobility, tunable band structure and different morphologies that are required for a good microwave absorbing performance.

The two-dimensional version of phosphorous like graphene is known as phosphorous or black phosphorus. This material unlike graphene has a band gap that can be tuned by changing the number of layers used (0.3 eV for its bulk version and 1.88 eV in a monolayer). The individual layers are held together by Van der Waals forces. Their monolayers show a structure of the quadrangular pyramid due to the bonds between their atoms with different angles (Hao, 2018).

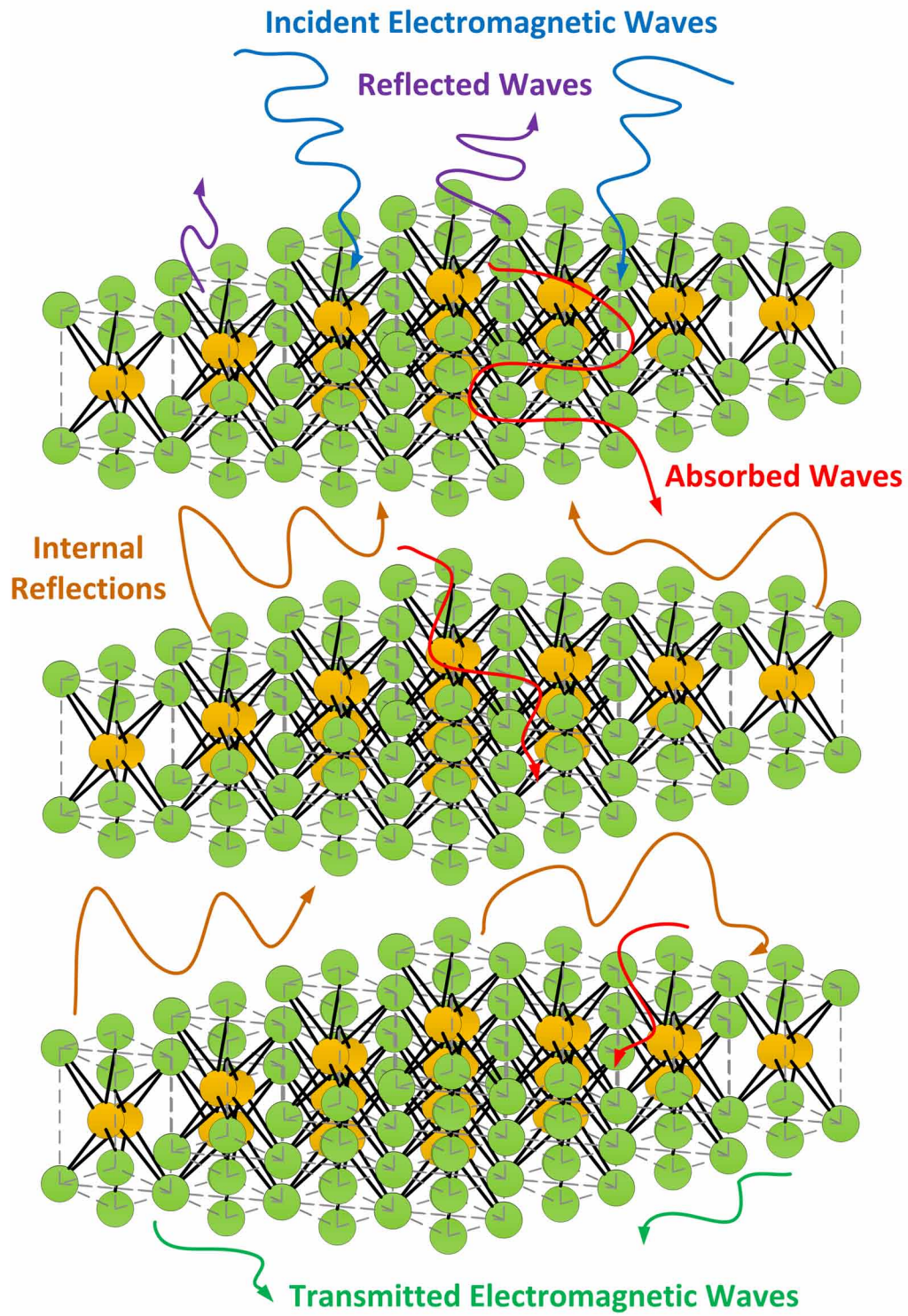
SOLUTIONS AND RECOMMENDATIONS

In this section, the chapter describes some practical examples where two-dimensional materials have been successfully applied in the shielding to electromagnetic interference of electromagnetic signals in various technological sectors. Two-dimensional materials (2D materials) are attractive for numerous applications due to their unusual physical and chemical properties derived from their strong interaction of electrons confined in a plane, and the effect of significant and anisotropic quantum confinement (Hu, 2018). In the cases of graphene and MXenes, their electromagnetic shielding properties are attributed to their special electrical behavior, large specific surface area, and low mass density. The inherent advantages of two-dimensional materials to be used as microwave attenuating materials are: 1) very large specific surface area and numerous interfaces for interfacial polarization as well as for the reflection and scattering of electromagnetic waves, 2) its ultra-small thickness that does not allow the penetration and dissipation of electromagnetic waves leads to multiple internal reflections and the dissipation of electromagnetic energy through topologies such as layer-by-layer assemblies, core-shell particles and porous foams, and 3) good processability that makes them easily integrable to multifunctional composites with multiple modes of attenuation of electromagnetic energy. In addition, these materials behave as microwave reflectors and absorbers, act as impedance regulators and provide structural support for good impedance matching and to establish the optimal structure (Hu, 2018). The different mechanisms present in two-dimensional materials to increase the shielding effectiveness (SE) are shown in Figure 2.

The first two-dimensional carbide (Ti₃C₂) was introduced in 2011 and from then until now numerous two-dimensional carbides, nitrides and carbonitrides called MXenes have been discovered or isolated, so that their technological development and scientific research are still in their infancy (Li, 2018). Microwave attenuation capacity was demonstrated for Ti₃C₂ nanosheet and paraffin-based composites in the 2 to 18 GHz range with a thickness of 2 mm presenting a reflection loss of 40 dB (Feng, 2018). The absorption mechanism is due to the high dielectric loss and strong multiple internal reflections among layers of the MXene.

Ti₃C₂T_x films with thickness in the range of 1 to 45 microns showed excellent electrical conductivity and electromagnetic interference shielding (Shahzad, 2016). The 45-micron thick film showed an effectiveness of 92 dB (more than 50 dB for a 2.5-micron film). According to the authors of the article, these materials present the highest effectiveness values of any known synthetic material of similar thickness. Its performance is derived from excellent electrical conductivity (4600 Siemens per centimeter) and multiple internal reflections in the case of free-standing films. The main electromagnetic shielding mechanism for the composite material was absorption instead of reflection in the case of pristine MXene. In this study they analyzed the electromagnetic shielding performance of MXene films based on Ti₃C₂T_x, Mo₂TiC₂T_x and Mo₂Ti₂C₃T_x. The shielding capacity is maintained (57 dB) even when the MXene is added to sodium alginate (SA) to create polymer composite films. SA is a linear polysaccharide copolymer

Figure 2. Mechanisms for electromagnetic interference shielding presented in two-dimensional materials



derived from marine algae that is abundant biomaterial, non-harmful to the environment, mechanically robust and with functional groups containing oxygen (OH^- , COO^- , and O^{2-}) that facilitate the formation of hydrogen bonds with functional groups of MXenes.

Composite materials based on cellulose nanofiber (CNF) and extremely thin and highly flexible $\text{Ti}_3\text{C}_2\text{T}_x$ MXene with a nacre-like lamellar structure were manufactured using a self-assembly process induced by vacuum filtration (Cao, 2018). For a film with a thickness of 47-microns, high tensile strength (up to 135.4 MPa), fracture strain (up to 16.7%), folding endurance (up to 14 260 times), high electrical conductivity (up to 739.4 S/m) and efficiency to the excellent specific electromagnetic shielding (up to 2647 dBcm²/g). In these composite materials, the main shielding mechanism is electromagnetic absorption and in the case of pristine MXene is electromagnetic reflection. The function of cellulose nanofiber is to provide high mechanical strength, great flexibility and reduce the number of insulating contacts between the conductive nanosheets of the two-dimensional material. This composite material can develop shielding for electromagnetic interference in applications such as flexible wearable devices, weapon equipment, and robot junctions.

Electrostatic assemblies using negative $\text{Ti}_3\text{C}_2\text{T}_x$ MXene nanosheets on positive polystyrene microspheres using compression molding with a low percolation threshold of 0.26 vol% presented an electrical conductivity of 1081 S/m and a shielding effectiveness greater than 54 dB or for a 1.90 vol% with an effectiveness of 62 dB were implemented (Sun, 2017a). Pristine MXene presents electromagnetic shielding due to reflection, while the composite material presents shielding due to electromagnetic absorption. The use of conductive polymer nanocomposites is favored by its excellent processability and its versatility of tuning properties over wide ranges. Their properties depend on their intrinsic conductivity, aspect ratio, dispersion, and distribution of conductive fillers. An excessive increase in the amount of filler used in the composite material is not feasible because its processing is difficult, and a catastrophic deterioration in its mechanical properties as coating occurs.

In the search for multifunctional materials, a material formed by the combination of Ti_3SiC_2 MXene and the composite formed between silicon carbide and carbon fibers (SiC/C) has been synthesized (Fan, 2018). Composite SiC/C for thermostructural applications has low density, high thermal shock resistance, good thermal stability, and excellent mechanical properties at high temperatures. Ti_3SiC_2 presents a high electrical conductivity and a shielding effectiveness (SE) against electromagnetic interference of 35 to 54 dB in the frequency range of 8.2 to 18 GHz (X-Band). The electromagnetic shielding mechanism for these materials is absorption, instead of reflection, which is presented in pristine MXene and composite SiC/C.

Most composites based on polymers and two-dimensional materials have a “brick and mortar” structure. Composites based on $\text{Ti}_3\text{C}_2\text{T}_x$ and poly(3,4-ethylenedioxythiophene)-poly(styrenesulfonate) (PEDOT-PSS) with a weight ratio of 7:1 with a thickness of 11.1 μm showed a shielding effectiveness of 42.1 dB (Liu, 2018). In addition, this film showed lightweight, a tensile strength of 13.71 MPa, an electrical conductivity of 340.5 S/cm and a specific shielding efficiency of 19,497.8 dBcm²/g. The main electromagnetic shielding mechanism in this material was absorption, instead of reflection, that it is presented in pristine MXene.

Composite materials based on $\text{Ti}_3\text{C}_2\text{T}_x$ MXene and non-woven carbon fabric offer lightweight and flexible materials with shielding effectiveness of 43.2 dB and specific shielding efficiency of 8534.7 dBcm²/g at 2.3 GHz (Raagulan, 2018). The main applications where these materials can be used are aviation, portable electronics, radars, aerospace, and military. The mechanism associated with electromagnetic interference shielding in this case is associated with both electromagnetic reflection and absorption.

Hybrid aerogels based on two-dimensional materials have been proposed as three-dimensional materials for electromagnetic interference shielding (Zhao, 2018b). The proposed two-dimensional materials

were $\text{Ti}_3\text{C}_2\text{T}_x$ MXene and reduced graphene oxide (rGO). The resulting material had an electrical conductivity of 695.9 S/m and a shielding effectiveness of more than 50 dB in the X-band, with a content of 0.74 vol% of MXene. The main electromagnetic shielding mechanism for the composite material was absorption, instead of reflection, in the case of the pristine MXene.

$\text{Ti}_3\text{C}_2\text{T}_x/\text{Ni}_{0.5}\text{Zn}_{0.5}\text{Fe}_2\text{O}_4$ composites obtained by in situ co-precipitation with a 5 wt% $\text{Ti}_3\text{C}_2\text{T}_x$ presented a shielding effectiveness of 42.5 dB in the K-Band (12–15 GHz) (Li, 2017). The electromagnetic wave absorption mechanism is complex, since it involves magnetic loss, dielectric loss, conductivity loss, multiple reflections, and scattering. These sub-mechanisms are difficult to achieve through a single material; therefore, hybrid materials provide an option of technological development to increase or guarantee a true shield against electromagnetic interference.

Composite materials based on molybdenum disulfide (MoS_2) nanostructures with a diameter of 40 to 60 nm and carbonyl iron@ SiO_2 particles were synthesized to form a microwave absorbing material (Wang, 2018). Materials of 1.6 mm thickness with 15 wt% of MoS_2 showed a minimum reflection loss of -28 dB at 12.4 GHz. This type of materials, in addition to having an easy synthesis process, presents a good absorption performance with thin thickness and mechanical strength at high temperatures. Novel materials based on molybdenum disulfide (MoS_2) have been designed using different reinforcements: 1) nickel nanoparticles, 2) carbon nanotubes using a one-dimensional architecture, and 3) carbon nanotubes using layers of three-dimensional carbon nanotubes (Sun, 2017b). For this group of materials, their shielding effectiveness (SE) were 19.7 dB at 2.92 GHz, 47.9 dB at 5.60 GHz, and 69.2 dB at 4.88 GHz, respectively. The control of interfacial contacts in the materials produced makes van der Waals heterostructures an alternative strategy to develop microwave attenuation properties. The main mechanism of electromagnetic shielding involved in reducing interference for this type of materials is based on electromagnetic absorption.

Electrically conductive networks based on hybrid materials formed by the stacking of layers of reduced graphene oxide (rGO) and molybdenum disulfide (MoS_2) (20 wt%) present electromagnetic absorption with a shielding effectiveness of 10 to 60 dB at a thickness of 2.5 mm (Xie, 2016). The main mechanism of electromagnetic shielding is electromagnetic absorption. Nanocomposites based on molybdenum disulfide (MoS_2), reduced graphene oxide (rGO) and magnetite nanoparticles (Fe_3O_4) were synthesized by hydrothermal process for electromagnetic shielding applications (Prasad, 2018). A shielding effectiveness of 8.27 dB, improved magnetic loss and matching impedance were achieved for the range of 8 to 12 GHz, thanks to dielectric loss by molybdenum disulfide and graphene, as well as magnetic loss by magnetite. In this way, the combination of material forming hybrid materials can provide better electromagnetic interference shielding.

An ultra-lightweight composite aerogel based on reduced graphene oxide (rGO) and black phosphorus (BP) or phosphorene was manufactured using a self-assembly process (Hao, 2018). This material exhibited a reflection loss of 46.9 dB for a thickness of 2.53 mm in a wide absorption band of 6.1 GHz, and which can be used as an ultralight microwave absorber. The shielding mechanism is mainly dominated by electromagnetic absorption produced by multiple internal reflections within the hybrid material. The structure between these two-dimensional materials formed four kinds of interconnected interfaces such as wrapping, wearing, bridging and weak linking.

Flexible polydimethylsiloxane (PDMS)-graphene composite foam using a chemical vapor deposition technique achieved a specific shielding effectiveness of 333 dBcm^3/g with 0.8 wt% graphene loading in the X-band frequency range (Dhakate, 2015). The effectiveness value is higher than that obtained for solid copper of 10 dBcm^3/g or 33.1 dBcm^3/g para PDMS-carbon nanotube composites with 7 wt% of carbon nanotubes. In this composite material, electromagnetic shielding is due to electromagnetic

absorption and multiple internal reflections produced within the foam. These types of materials can be used in applications such as aircraft, spacecraft, automobiles and portable electronics. Composites based on poly(dimethylsiloxane) and graphene were synthesized using 3.07 wt% of graphene and presented an electrical conductivity of 103 S/m and shield effectiveness of 54 dB in the X-band (Xu, 2018). These materials behave flexible, so they can be used in wearable devices and as shelters to protect against electromagnetic interference. Another option for composite materials for electromagnetic shielding based on polydimethylsiloxane (PDMS) was reinforced by reduced graphene oxide (rGO) and single-walled carbon nanotubes (SWCNTs) (Zhao, 2018a). These materials reached, in addition to being flexible and having lightweight, an electrical conductivity of 1.2 S/cm and a shielding effectiveness of 31 dB over the range of the X-band.

Hybrid materials have also been synthesized to improve the shielding effectiveness based on silicone rubber vinyl methyl silicone (VMQ), graphene nanoplate, magnetite (Fe_3O_4) @ barium titanate (BaTiO_3) (GMBT) nanoparticles and multiple wall nanotubes (MWCNTs) (Jin, 2018). The shielding effectiveness for this material was in the range of 26.7 to 33.3 dB (attenuation greater than 99.8%) with a load of 16.1 wt% total filler (GMBT: MWCNTs = 5:1). The shielding mechanism is mainly dominated by the absorption of electromagnetic radiation, thanks to the multiple internal reflections within the hybrid material.

Hybrid nanocomposites based on poly(methyl methacrylate), graphene and magnetite (Fe_3O_4) have been used as shielding material for electromagnetic interference (Sharif, 2017). Nanocomposites with a 0.3 vol% rGO obtained an electrical conductivity of 91.2 S/m and presented a shielding effectiveness of 63.2 dB for a 2.9 mm film. In this composite material, the main mechanism for allowing electromagnetic shielding is the electromagnetic absorption produced by the multiple internal reflections presented.

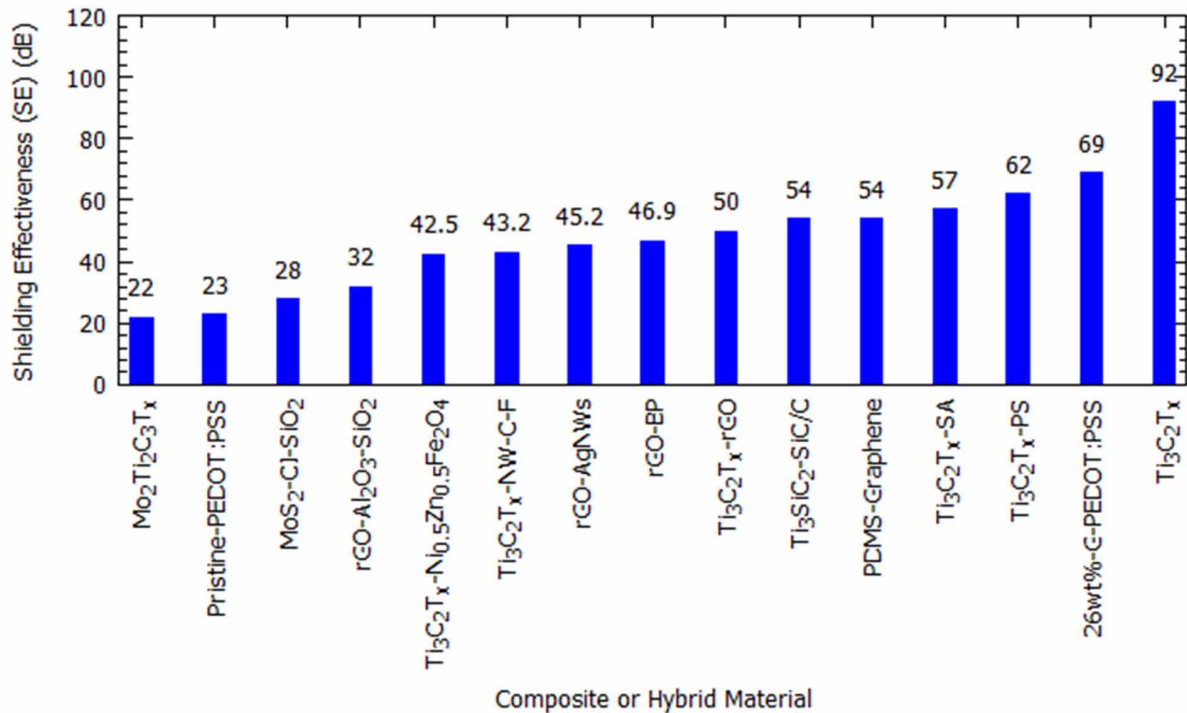
Aerogels based on reduced graphene oxide (rGO) and silver nanowires (AGNWs) were synthesized, and these reaching a shielding effectiveness of 45.2 dB and a mechanical strength of 42 KPa (Liu, 2019). They are ideal for applications that require mechanical robustness and shielding against electromagnetic interference. The shielding mechanism is mainly dominated by the absorption of electromagnetic radiation, thanks to the multiple internal reflections within the hybrid material. In this structure, the one-dimensional silver nanowires attached to the graphene sheets act as a bridge among the sheets to facilitate the formation of an electron-hopping conductive network.

Electromagnetic shielding for harsh environments requires the use of ceramic composites with good electrical conductivity, lightweight, and high mechanical strength. A composite based on mullite and reduced graphene oxide (rGO) has been sintered by plasma at 1200 °C using $\gamma\text{-Al}_2\text{O}_3\text{@SiO}_2$ powder as a precursor (Ru, 2018). This material achieved a shielding effectiveness greater than 32 dB and an electrical conductivity of 696 S/m using only 0.89 vol% rGO to operate in the K-band without losing its mechanical properties. The mechanism of multiple internal reflections is the main responsible for electromagnetic shielding. This material can be used in applications where high mechanical stresses and high temperatures are presented.

Several types of dielectric spacers based on graphene, graphene oxide and polymers that result in materials with electrical loss and magnetic loss have provided improvements in shielding effectiveness (SE) (Song, 2017). These materials presented a shielding effectiveness higher than 20 dB in the X-band thanks to the frequency selectivity associated with them. The main mechanism of electromagnetic shielding is based on multiple internal reflections.

Sandwiches formed from stacked films with multiple layers based on: 1) graphene oxide (GO) doped with nitrogen, and 2) copper phthalocyanine, have been synthesized as materials for electromagnetic shielding (Wang, 2017). These materials present a shielding effectiveness of up to 55.2 dB for a thick-

Figure 3. Shielding effectiveness for different two-dimensional materials



ness of 0.47 mm and an electrical conductivity of 3.64×10^3 S/m. The mechanism of absorption is the main responsible for electromagnetic shielding.

A comparison of the performance of different types of materials based on two-dimensional materials to improve shielding effectiveness is illustrated in Figure 3.

FUTURE RESEARCH DIRECTIONS

The introduction of two-dimensional materials not only seeks to effectively shield electromagnetic interference but also to reduce the weight of cars and airplanes to increase energy efficiency. Future materials used in the shielding of electromagnetic interference should exploit the quantum confinement of electrons (restrict the movement of electromagnetic waves), extremely small thickness (increase the number of multiple reflections and dissipation of electromagnetic energy), the specific surface area large (improved interfacial polarization for reflection and dispersion of electromagnetic waves), and low mass density (reduce the vibration of the material by electromagnetic waves) offered by two-dimensional materials. In addition, two-dimensional materials should continue to exploit the layer-by-layer assemblies, core-shell structures and porous foams to increase the multimode attenuations of electromagnetic waves so that an electrical impedance matching occurs, which ensures electro-magnetic shielding to a wider range of electrical frequencies.

Graphene has the lowest shielding effectiveness between two-dimensional materials, but it can be exploited more efficiently in composite materials or hybrid materials. Scientific research on lightweight composite foams using polymers and graphene should promote future developments to further increase

the efficiency of electromagnetic shielding. Technological development and scientific research involved in the synthesis of composite or hybrid materials exploiting graphene as a material to increase shielding effectiveness should be developed for at least a decade. Despite the advances achieved so far, significant challenges must be overcome to achieve electrical conductivity and effectiveness to specific shielding effectiveness with desirable thicknesses using composite and hybrid materials based on two-dimensional materials. Scientific research and technological development related to the use of MXenes in synthesis of materials for electromagnetic shielding should be carried out by at least two decades. Unprecedented efforts must be made to theoretically predict the behavior of two-dimensional materials with respect to the shielding effectiveness against electromagnetic interference since each material has very different properties and structures.

CONCLUSION

In this chapter, the advantages of using two-dimensional materials such as graphene, transition metal dichalcogenides, MXenes and phosphorene as absorbent materials to electromagnetic waves as well as materials to increase the shielding effectiveness to electromagnetic interference in various areas of application were presented. The possibility of using hybrid and composite materials to improve the electrical and magnetic properties and thereby favor their application in electromagnetic interference shielding is illustrated through different examples with a wide variety of feasible combinations of materials. These materials can be applied in fields such as electromagnetic pollution, radar wave absorption, satellite communication, wireless telecommunications, in the care of human health, protection of electronic devices, aeronautical systems, space systems, and military systems.

REFERENCES

- Anasori, B., & Gogotsi, Y. (Eds.). (2019). *2D Metal Carbides and Nitrides (MXenes): Structure, Properties and Applications*. Springer. doi:10.1007/978-3-030-19026-2
- Bhattacharjee, Y., Arief, I., & Bose, S. (2017). Recent Trends in Multilayered Architectures towards Screening Electromagnetic Radiation: Challenges and Perspectives. *Journal of Materials Chemistry. C, Materials for Optical and Electronic Devices*, 5(30), 7390–7403. doi:10.1039/C7TC02172K
- Cao, M.-S., Shu, J.-C., Wang, X.-X., Wang, X., Zhang, M., Yang, H.-J., Fang, X.-Y., & Yuan, J. (2019). Electronic Structure and Electromagnetic Properties for 2D Electromagnetic Functional Materials in Gigahertz Frequency. *Annalen der Physik*, 531(4), 1800390. doi:10.1002/andp.201800390
- Cao, W.-T., Chen, F.-F., Zhu, Y.-J., Zhang, Y.-G., Jiang, Y.-Y., Ma, M.-G., & Chen, F. (2018). Binary Strengthening and Toughening of MXene/Cellulose Nanofiber Composite Paper with Nacre-Inspired Structure and Superior Electromagnetic Interference Shielding Properties. *ACS Nano*, 12(5), 4583–4593. doi:10.1021/acsnano.8b00997 PMID:29709183
- Dhakate, S. R., Subhedar, K. M., & Singh, B. P. (2015). Polymer Nanocomposite Foam Filled with Carbon Nanomaterials as an Efficient Electromagnetic Interference Shielding Material. *RSC Advances*, 5(54), 43036–43057. doi:10.1039/C5RA03409D

- Fan, X., Yin, X., Cai, Y., Zhang, L., & Cheng, L. (2018). Mechanical and Electromagnetic Interference Shielding Behavior of C/SiC Composite containing Ti_3SiC_2 . *Advanced Engineering Materials*, 20(2), 1700590. doi:10.1002/adem.201700590
- Feng, W., Luo, H., Wang, Y., Zeng, S., Deng, L., Zhou, X., Zhang, H., & Peng, S. (2018). Ti_3C_2 MXene: A Promising Microwave Absorbing Material. *RSC Advances*, 8(5), 2398–2403. doi:10.1039/C7RA12616F
- Hao, C., Wang, B., Wen, F., Mu, C., Xiang, J., Li, L., & Liu, Z. (2018). Superior Microwave Absorption Properties of Ultralight Reduced Graphene Oxide/Black Phosphorus Aerogel. *Nanotechnology*, 29(23), 235604. doi:10.1088/1361-6528/aab83a PMID:29558364
- Hu, M., Zhang, N., Shan, G., Gao, J., Liu, J., & Li, R. K. Y. (2018). Two-Dimensional Materials: Emerging Toolkit for Construction of Ultrathin High-Efficiency Microwave Shield and Absorber. *Frontiers in Physics*, 13(4), 138113. doi:10.1007/11467-018-0809-8
- Jin, L., Zhao, X., Xu, J., Luo, Y., Chen, D., & Chen, G. (2018). The Synergistic Effect of a Graphene Nanoplate/ Fe_3O_4 @ BaTiO_3 Hybrid and MWCNTs on Enhancing Broadband Electromagnetic Interference Shielding Performance. *RSC Advances*, 8(4), 2065–2071. doi:10.1039/C7RA12909B
- Kolobov, A. V., & Tominaga, J. (2016). *Two-Dimensional Transition-Metal Dichalcogenides*. Springer. doi:10.1007/978-3-319-31450-1
- Li, Y., Zhou, X., Wang, J., Deng, Q., Li, M., Du, S., Han, Y.-H., Lee, J., & Huang, Q. (2017). Facile Preparation of in situ Coated $\text{Ti}_3\text{C}_2\text{T}_x/\text{Ni}_{0.5}\text{Zn}_{0.5}\text{Fe}_2\text{O}_4$ Composites and Their Electromagnetic Performance. *RSC Advances*, 7(40), 24698–24708. doi:10.1039/C7RA03402D
- Li, Z., Wang, Z., Lu, W., & Hou, B. (2018). Theoretical Study of Electromagnetic Interference Shielding of 2D MXenes Films. *Metals*, 8(8), 652. doi:10.3390/met8080652
- Liu, R., Miao, M., Li, Y., Zhang, J., Cao, S., & Feng, X. (2018). Ultrathin Biomimetic Polymeric $\text{Ti}_3\text{C}_2\text{T}_x$ MXene Composite Films for Electromagnetic Interference Shielding. *ACS Applied Materials & Interfaces*, 10(51), 44787–44795. doi:10.1021/acsami.8b18347 PMID:30516359
- Liu, X., Chen, T., Liang, H., Qin, F., Yang, H., & Guo, X. (2019). Facile Approach for a Robust Graphene/Silver Nanowires Aerogel with High-Performance Electromagnetic Interference Shielding. *RSC Advances*, 9(1), 27–33. doi:10.1039/C8RA08738E
- Prasad, J., Singh, A. K., Shah, J., Kotnala, R. K., & Singh, K. (2018). Synthesis of MoS_2 -Reduced Graphene Oxide/ Fe_3O_4 Nanocomposite for Enhanced Electromagnetic Interference Shielding Effectiveness. *Materials Research Express*, 5(5), 055028. doi:10.1088/2053-1591/aac0c2
- Raagulan, K., Braveenth, R., Jang, H. J., Lee, Y. S., Yang, C.-M., Kim, B. M., Moon, J. J., & Chai, K. Y. (2018). Fabrication of Nonwetting Flexible Free-Standing MXene-Carbon Fabric for Electromagnetic Shielding in S-Band Region. *Bulletin of the Korean Chemical Society*, 39(12), 1412–1419. doi:10.1002/bkcs.11616
- Ru, J., Fan, Y., Zhou, W., Zhou, Z., Wang, T., Liu, R., Yang, J., Lu, X., Wang, J., Ji, C., Wang, L., & Jiang, W. (2018). Electrically Conductive and Mechanically Strong Graphene/Mullite Ceramic Composites for High-Performance Electromagnetic Interference Shielding. *ACS Applied Materials & Interfaces*, 10(45), 39245–39256. doi:10.1021/acsami.8b12933 PMID:30346124

Shahzad, F., Alhabeb, M., Hatter, C. B., Anasori, B., Hong, S. M., Koo, C. M., & Gogotsi, Y. (2016). Electromagnetic Interference Shielding with 2D Transition Metal Carbides (MXenes). *Science*, 353(6304), 1137–1140. doi:10.1126/science.aag2421 PMID:27609888

Sharif, F., Arjmand, M., Moud, A.A. & Sundararaj, U. (2017). Segregated Hybrid Poly(methyl methacrylate)/Graphene/Magnetite Nanocomposites for Electromagnetic Interference Shielding. *ACS Applied Materials & Interfaces*, 9(16), 14171–14179. DOI: doi:10.1021/acsami.6b13986

Sharon, M., & Sharon, M. (2015). *Graphene: An Introduction to the Fundamentals and Industrial Applications*. Wiley. doi:10.1002/9781118842577

Song, W.-L., Gong, C., Li, H., Cheng, X.-D., Chen, M., Yuan, X., Chen, H., Yang, Y., & Fang, D. (2017). Graphene-based Sandwich Structures for Frequency Selectable Electromagnetic Shielding. *ACS Applied Materials & Interfaces*, 9(41), 36119–36129. doi:10.1021/acsami.7b08229 PMID:28945066

Sun, R., Zhang, H.-B., Liu, J., Xie, X., Yang, R., Li, R., Hong, S., & Yu, Z.-Z. (2017a). Highly Conductive Transition Metal Carbide/Carbonitride-(MXene)@Polystyrene Nanocomposites Fabricated by Electrostatic Assembly for Highly Efficient Electromagnetic Interference Shielding. *Advanced Functional Materials*, 27(45), 1702807. doi:10.1002/adfm.201702807

Sun, Y., Zhong, W., Wang, Y., Xu, X., Wang, T., Wu, L., & Du, Y. (2017b). MoS₂-based Mixed-Dimensional van der Waals Heterostructures: A New Platform for Excellent and Controllable Microwave-Absorption Performance. *ACS Applied Materials & Interfaces*, 9(39), 34243–34255. doi:10.1021/acsami.7b10114 PMID:28901126

Wang, H., & Jin, P. (2018). The Electromagnetic and Microwave Absorbing Properties of MoS₂/Carbonyl Iron@SiO₂. *Materials Research Express*, 5(11), 116301. doi:10.1088/2053-1591/aadd80

Wang, Z., Wei, R., & Liu, X. (2017). Fluffy and Ordered Graphene Multilayer Films with Improved Electromagnetic Interference Shielding over X-Band. *ACS Applied Materials & Interfaces*, 9(27), 22408–22419. doi:10.1021/acsami.7b04008 PMID:28640583

Xie, A., Sun, M., Zhang, K., Jiang, W., Wu, F., & He, M. (2016). In Situ Growth of MoS₂ Nanosheets on Reduced Graphene Oxide (rGO) Surfaces: Interfacial Enhancement of Absorbing Performance against Electromagnetic Pollution. *Physical Chemistry Chemical Physics*, 18(36), 24931–24936. doi:10.1039/C6CP04600B PMID:27546736

Xu, F., Chen, R., Lin, Z., Qin, Y., Li, Y., Zhao, X., Yang, M., Sun, X., Wang, S., Peng, Q., Li, Y., & He, X. (2018). Superflexible Interconnected Graphene Network Nanocomposites for High-Performance Electromagnetic Interference Shielding. *ACS Omega*, 3(3), 3599–3607. doi:10.1021/acsomega.8b00432 PMID:31458611

Zhao, S., Yan, Y., Gao, A., Zhao, S., Cui, J., & Zhang, G. (2018a). Flexible Polydimethylsilane Nanocomposites enhanced with a Three-Dimensional Graphene/Carbon Nanotube Bicontinuous Framework for High-Performance Electromagnetic Interference Shielding. *ACS Applied Materials & Interfaces*, 10(31), 26723–26732. doi:10.1021/acsami.8b09275 PMID:29989792

Zhao, S., Zhang, H. B., Luo, J.-Q., Wang, Q.-W., Xu, B., & Yu, Z.-Z. (2018b). Highly Electrically Conductive Three-Dimensional $\text{Ti}_3\text{C}_2\text{T}_x$ MXene/Reduced Graphene Oxide Hybrid Aerogels with Excellent Electromagnetic Interference Shielding Performances. *ACS Nano*, 12(11), 11193–11202. doi:10.1021/acsnano.8b05739 PMID:30339357

ADDITIONAL READING

Gejo, G., Sanu, M. S., Prakashan, V. P., Sajna, M. S., Faisal, M., Wilson, R., Chandran, A., Biju, P. R., Cyriac, J., & Unikirshnan, N. V. (2018). Green and Facile Approach to Prepare Polypropylene/in situ Reduced Graphene Oxide Nanocomposites with Excellent Electromagnetic Interference Shielding Properties. *RSC Advances*, 8(53), 30412–30428. doi:10.1039/C8RA05007D

Hsiao, S.-T., Ma, C.-C. M., Tien, H.-W., Liao, W.-H., Wang, Y.-S., Li, S.-M., Yang, C.-Y., Lin, S.-C., & Yang, R.-B. (2015). Effect of Covalent Modification of Graphene Nanosheets on the Electrical Property and Electromagnetic Interference Shielding Performance of a Water-Borne Polyurethane Composite. *ACS Applied Materials & Interfaces*, 7(4), 2817–2826. doi:10.1021/am508069v PMID:25569714

Li, S., Tan, Y., Xue, J., Liu, T., Zhou, X., & Zhang, H. (2018). Electromagnetic Interference Shielding Performance of Nano-Layered Ti_3SiC_2 Ceramics at High-Temperatures. *AIP Advances*, 8(1), 015027. doi:10.1063/1.5012607

Li, Y., Shen, B., Yi, D., Zhang, L., Zhai, W., Wei, X., & Zheng, W. (2017). The Influence of Gradient and Sandwich Configurations on the Electromagnetic Interference Shielding Performance of Multilayered Thermoplastic Polyurethane/Graphene Composite Foams. *Composites Science and Technology*, 138, 209–216. doi:10.1016/j.compscitech.2016.12.002

Qing, Y., Wen, Q., Luo, F., Zhou, W., & Zhu, D. (2016). Graphene Nanosheets/ BaTiO_3 Ceramic as Highly Efficient Electromagnetic Interference Shielding Materials in the X-Band. *Journal of Materials Chemistry. C, Materials for Optical and Electronic Devices*, 4(2), 371–375. doi:10.1039/C5TC03035H

Ren, F., Song, D., Li, Z., Jia, L., Zhao, Y., Yan, D., & Ren, P. (2018). Synergistic Effect of Graphene Nanosheets and Carbonyl Iron-Nickel Alloy Hybrid Filler on Electromagnetic Interference Shielding and Thermal Conductivity of Cyanate Ester Composites (2018). *Journal of Materials Chemistry. C, Materials for Optical and Electronic Devices*, 6(6), 1476–1486. doi:10.1039/C7TC05213H

Tan, Y. Q., Luo, H., Zhou, X. S., Peng, S. M., & Zhang, H. B. (2018). Boron Carbide Composites with Highly Aligned Graphene Nanoplatelets: Light-Weight and Efficient Electromagnetic Interference Shielding Materials at High Temperatures. *RSC Advances*, 8(69), 39314–39320. doi:10.1039/C8RA07351A

Yuan, Y., Yin, W., Yang, M., Xu, F., Zhao, X., Li, J., Peng, Q., He, X., Du, S., & Li, Y. (2018). Light-weight, Flexible and Strong Core-Shell Non-Woven Fabrics covered by Reduced Graphene Oxide for High-Performance Electromagnetic Interference Shielding. *Carbon*, 130, 59–68. doi:10.1016/j.carbon.2017.12.122

KEY TERMS AND DEFINITIONS

Black Phosphorus (BP): Stable orthorhombic form of phosphorus with two-dimensional structure at room temperature and pressure of entangled rings with six members where each atom is linked to three other atoms.

Electromagnetic Interference (EMI) Shielding: Conductive or magnetic barrier material used to reduce the effect of the electromagnetic field in a space by blocking its effect.

Graphene: Allotrope of carbon formed by a single layer of carbon atoms located in a hexagonal lattice.

Mxene: Two-dimensional inorganic compound with thickness layers of few atoms based on transition metal carbides, nitrides, or carbonitrides.

Shielding Effectiveness: The product of 20 times the logarithm of the ratio between the incident emission and the transferred emission given in decibels (dB). A shielding effectiveness of 40 dB provides 99 percent attenuation of electromagnetic radiation that represents the minimum required for most applications.

Transition Metal Dichalcogenide: Chemical compound consisting of two chalcogenic anions (sulfur, selenium, tellurium) and at least one transition metal (molybdenum, tungsten, cobalt, etc.) with two-dimensional structure.

Two-Dimensional Material: Crystalline material consisting of a single layer of atoms or a single layer of two or more covalently bonding elements.

Effect of Nanoparticle and Inclination Angle at Thermal Efficiency in Heatpipes

Sina Razvarz

 <https://orcid.org/0000-0003-1549-7307>

Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico

Raheleh Jafari

 <https://orcid.org/0000-0001-7298-2363>

School of Design, University of Leeds, Leeds, UK

Cristóbal Vargas Jarillo

Department of Automatic Control, Centre for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico

Alexander Gegov

 <https://orcid.org/0000-0002-6166-296X>

School of Computing, University of Portsmouth, UK

INTRODUCTION

Recently, the proficiency of the engineering applications have become more notable. Continuous advancement have become essential in order to maintain the technology itself. Technology has improved significantly by accepting novel opinions and approaches for effective application, storage as well as transmission of energy. Nanotechnology is one of the most noteworthy fields which fascinates the researchers because of its superior capability, enabling a considerable advancement in the efficiency of several devices. Over the past few decades remarkable investigations have been carried out in order to implement nanotechnology in heat transfer usages. Nanofluids can be considered as a novel improvement in thermo-fluidics, acquired by stirring nano particles in conventional fluids (Choi, 1998). Investigators are taken nanoparticles of various metals, metal oxides, carbides, nitrides, as well as various kinds of carbon with several base fluids to be mentioned as water, ethylene glycol (EG) as well as engine oils.

Industrial engineering is one of the fields that artificial neural network, fuzzy logic, and nano technology have found an extensive implementation area (Razvarz et al., 2017; Jafari & Yu, 2018 ; Jafari & Razvarz, 2017, 2018, Jafari et al., 2016, 2018a, 2018b, 2019a, 2019b; Jafarian et al., 2012). Nanoscale refers size dimensions between approximately 1 and 100 nm (or more appropriately, 0.2 and 100 nm) because at this scale the properties of materials differ with respect to their physical properties from a larger scale.

The size dimensions of Nanoscale are among nearly 1 and 100 nm (or more suitably, 0.2 and 100 nm), since at this scale the attributes of materials change concerning their physical attributes from a greater scale. Every material which contains one or more dimensions in the nanoscale is stated as nanomaterial (Sekhon, 2014). Nanoparticles contain distinct biological attributes such as tiny size, a great surface-area-to-volume.

DOI: 10.4018/978-1-7998-3479-3.ch087

Nanofluids have been constructed by utilizing two techniques: one step technique and two step technique (Eastaman et al., 2001). The principal aim of these techniques is to provide a homogeneous as well as stable solid-liquid mixture, also avoiding the agglomeration, feasible erosion and clogging etc. In one step technique, the nanoparticles have been constructed and dispersed into the foundation fluid at the same time. Whereas in a two step technique, the nanoparticle has been generated in the first step and dispersed into host fluids in the second step. Investigation reveals that nanofluids with oxide nanoparticles and carbon nanotubes can be produced utilizing two step technique. Two step technique is not appropriate for metallic nanoparticles. Under particular circumstances, nanofluids represent excellent thermo physical features that result in better effect of the thermal usage (Singh & Raykar, 2008).

Currently, so many investigators have suggested mechanisms of increment in heat transfer via nanofluid as the affect of Interface, Brownian motion, Ballistic conveyor of energy carriers as well as thermophoresis. The first attempt in adding nano materials or particles in the liquid was reported by Choi in (Choi et al., 1999; Choi, 1995; Pinto & Fiorelli, 2019) which was named nanofluid in 1995. It is a novel class of heat transfer as well as cooling working fluid. In (Kebllinski et al., 2005) the authors analyzed the consequence of adding nanoparticles to liquid which improved defects of past that the adding of millimeter as well as micro meter scale particles may lead to the sedimentation and pipe block. The heat transfer specifications of nanofluid are discussed in (Lin et al., 2008). The mechanisms that are responsible for the increment in thermal conductivity of nanofluids are investigated in (Chandrasekar et al., 2010; Ren, 2019). Furthermore, several mathematical models for identifying various factors as well as their limitations have been investigated in recent years. In (Murshed et al., 2005; Wang et al., 2017) the authors have been proposed that the nanoparticles shape, size as well as volume fraction are the principal parameters for deciding the thermal conduction of nanofluids. The applications as well as the challenges of nanofluids have been studied in (Saidur et al., 2011).

Heat pipe is considered as a device utilized in conduction the heat from one place to another. They are passive systems which convey heat from a heat origin (evaporator) to a heat sink (condenser) over lengthy intervals through the hidden heat of vaporization of a working fluid. The heat pipe can be theoretically classified into three parts, the evaporator, the adiabatic section, as well as the condenser. A bare heat pipe structure is demonstrated in Figure 1. In procedure, the working fluid attracts the heat from the heat source, then evaporates in the evaporator. The warm vapor straightly runs via the adiabatic part, and then condenses in the condenser. The condensed working fluid in the condenser runs back through capillary construction on the internal wall in the adiabatic part and then attains the evaporator once more. The idea of the heat pipe was initially represented in (Gaugler, 1944; Trefethen, 1962; Rakesh & Bumatraria, 2019), however, was not extensively published up to an independent growth by (Grover et al., 1964) at the Los Alamos Scientific Laboratories. Thermosyphon heat pipe is an empty, closed, and evacuated pipe which contains a working fluid. It is partitioned into three parts: evaporator, adiabatic, and the condenser part. The working fluid attracts the heat in evaporator part, and then gets transformed into the vapor phase that moves to the condenser part because of the pressure variance, declines heat to the cooling media via the condenser part and consequently gets transformed into the liquid phase.

BACKGROUND

The main constituents of a heat pipe are a sealed container, a wick construction, as well as a working fluid. All these components are working simultaneously in order to convey the heat more effectively and evenly. The wick construction is positioned on the internal area of the heat pipe wall, also is saturated

with the liquid working and makes the construction to produce the capillary action for liquid regression from the condenser to the evaporator. Regression of the liquid to the evaporator from the condenser is produced by wick construction.

Heat Pipe Container Sealed Containment Shell or Vessel

The container, working fluid, as well as wick construction of a heat pipe specify its operational specifications. One of the most significant factor in selecting the material for the heat pipe container as well as the wick is its adaptability with the working fluid. Decay of the container or wick as well as the pollution of the working fluid because of the chemical reaction may critically weaken heat pipe efficiency. For instance, the noncondensable gas generated during a chemical reaction ultimately may stack near the end of the condenser, reducing the condensation surface region (Naveen Kumar Gupta, 2019). This decreases the capability of the heat pipe to transfer heat to the exterior heat sink. Furthermore, the material as well as the geometry of the heat pipe container should have a supreme burst power, low weight, supreme thermal conductivity, as well as low porosity.

Because of the gravity, the condensed working fluid comes back into the evaporator part. So, the heat is attracted continuously via the evaporator part also is freed by the condenser part. There exist several kinds of heat pipes, which are extensively utilized in various usages to be mentioned as micro grooved, mesh wick, sintered wick, oscillating, and thermosyphon etc.

Working Fluid

The wick generates the construction for developing the capillary action for the liquid comes back from the condenser (heat output/sink) to the evaporator (heat input/source). As the heat pipe has a vacuum, the working fluid boils and takes up the hidden heat at well below its boiling point at atmospheric pressure. For example, the water boils at only over 273°K (0°C) and starts to efficiently convey hidden heat at this down temperature. The selection of the heat pipe working fluid is based on the operating temperature range of the usage. Working fluids range from liquid helium for exceedingly down temperature usages (-271°C) to silver ($>2,000^{\circ}\text{C}$) for exceedingly supreme temperatures. Water can be considered as one of the most usual heat pipe working fluid for an operating temperature range from 1°C to 325°C . Down temperature heat pipes utilize fluids like ammonia as well as nitrogen. High temperature heat pipes use cesium, potassium, NaK as well as sodium ($873\text{--}1,473^{\circ}\text{K}$). the range of working of some fluids are presented in Table 1.

In experimental/theoretical researches, investigators have utilized various kinds of nanofluids to be mentioned as Aluminium oxide (Do et al., 2010a; 2010b; Putra et al., 2012, 2014; Noie et al., 2009; Kamyar et al., 2013; Heris, 2014; Ji et al., 2013; Qu et al., 2010, Hung et al., 2012; Mashaei et al., 2016) Copper oxide (Wang et al., 2010, 2017, Liu et al., 2010, 2007, 2013; Liu & Zhu 2011; Venkatachalapathy et al., 2015; Putra et al., 2014; Chen et al., 2013; Alizad et al., 2012), Copper (Liu et al., 2011; Kole & Dey, 2013; Sokomon et al., 2012; Senthikumar et al., 2012; Riehl & Santos, 2012; Wan et al., 2015; Chen et al., 2013; Solomon et al., 2012, 2014) Magnesium oxide (Menlik, 2014), ferric oxide (Liu et al., 2013 Goshayeshi et al., 2015, 2016; Huminic & Huminic, 2013), Silicon dioxide (SiO_2) (Gunnasegaran et al., 2015, 2013; Qu & Wu, 2011; Chen et al., 2013), Titanium dioxide (Shafahi et al., 2010; Alizad et al., 2012) and Zinc oxide (Saleh et al., 2013) etc. in order to acquire the increased thermal performance. Thermal performance contains thermal effectiveness, thermal resistance, the efficient thermal conduc-

Table 1. Temperature Range of working fluid

Working Fluid	Operating Temperature Range (°K)
Hydrogen	13-43
Carbon Dioxide	223.15-303.15
Methane	93-173
Oxygen	63-143
Neon	33-43
Nitrogen	73-113
Helium	2- 4
Acetone	225-398
Ethane	123-298
Ammonia	198-398
Methanol	198-393
Pentane	148-398
Methylamine	183-398
Water	274-598
Caesium	623-1198
Sodium	773-1498
Lithium	1198-2098
Silver	1898-2298
Propylene	123-333
Potassium	673-1298

tivity of heat pipe, surface temperature gradient, convective heat transfer coefficient at evaporator and condenser part etc.

Wick Structure

The heat pipe wick construction utilizes capillaries for removing the liquid working fluid from condenser back to the evaporator part. The heat pipe wick construction is made of different materials and approaches. The most popular heat pipe wick construction has axially grooves on the internal heat pipe vessel wall. The wick construction as well as the working fluid produce the capillary forces needed to pump liquid from the condenser to the evaporator, also hold liquid equally spread in the wicking material. The heat pipe wicks are categorized into homogeneous wicks and composite wicks. Homogeneous wick is made of a single material as well as configuration. The most well-known kind of homogeneous wicks are wrapped screen, sintered metal as well as axial groove. Composite wick is made of two or more materials and configurations. The most popular kinds of composite wicks are variable screen mesh, screen-covered groove, screen slab accompanied by grooves, as well as screen tunnel accompanied by grooves. However, the required material features as well as the constructional properties of heat pipe wick constructions are a supreme thermal conductivity, supreme wick porosity, tiny capillary radius, and supreme wick permeability (Ghanbarpour et al., 2015a, 2015b; Senthil et al., 2016).

Figure 1. wicks contain single materials (left) wicks contain two or more materials(Right)



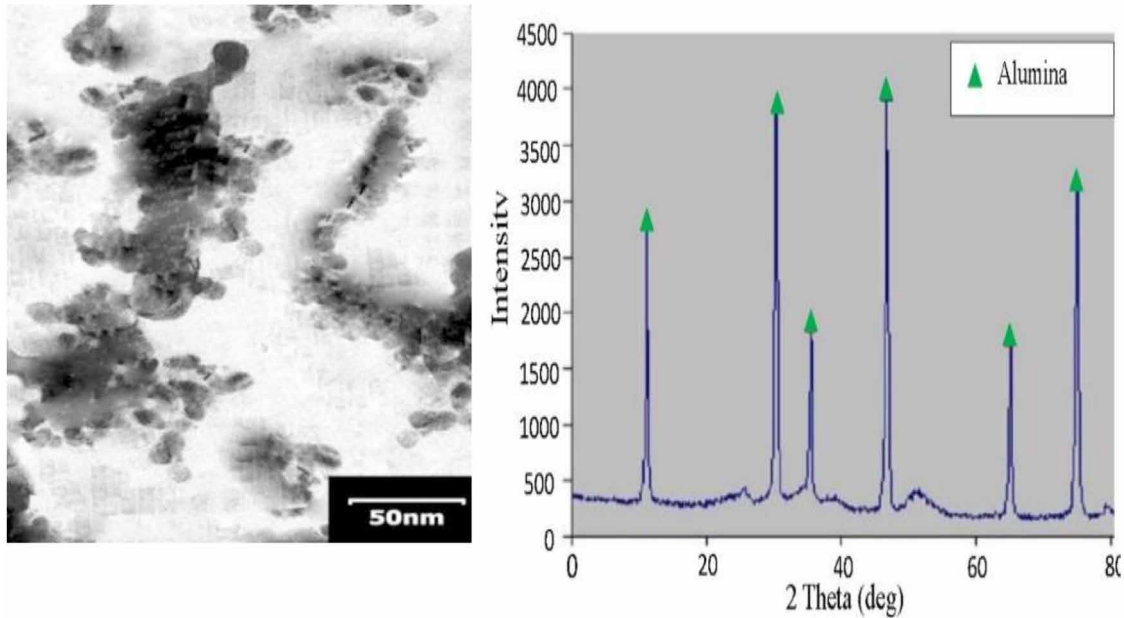
The wick helps the liquid to flow from the condenser to the evaporator part of the heat pipe. Furthermore, it supplies surface pores, which are essential at the liquid–vapor interface in order to develop the needed capillary pressure. Besides, the wick construction effects the radial temperature fall at the evaporator end among the internal heat pipe surface as well as the liquid–vapor surface. Hence, an efficient wick needs great inner pores in a direction normal to the heat flow way. This will cause a minimization in the liquid flow resistance. Moreover, tiny surface pores are needed in order to develop a supreme capillary pressure as well as a great conductive heat flow direction in order to minimize the radial surface to liquid–vapor surface temperature fall. For satisfying these necessities, two kinds of wick construction are generated. These are the homogeneous wicks which are constructed of a single material, and the composite wicks contain two or more materials, see Figure 1.

EXPERIMENTAL SETUP AND APPROACH

Characterization and Preparation of Aluminium Oxide Nanofluids

The fluid used in this experiment is pure water, also Aluminum oxide nanoparticles are the most popular nanoparticles utilized with 30 nm as well as 45 nm in size evaluated by transmission electron microscopy (TEM). The TEM photograph and XRD pattern of Aluminum oxide nanoparticles along with particle size around 35 nm added to pure water is demonstrated in Figure. 2. The nanofluid utilized in this experiment is constructed of 99.0% pure Aluminum oxide along with an average particle size of 35nm pre-dispersed in water also the precursors for its thermal deposition are trimethylaluminum (TMA) and water. Moreover, XRD is utilized in order to trace the transformation of Aluminum oxide. It is clear from this Figure that Aluminum oxide is observed in the as-sprayed nano particles. Aluminum oxide particles true density is considered to be 3880 kg/m^3 that is able to convert into weight fraction as well as volume fraction. The nanoparticles as well as the pure water are combined by using an ultrasonic homogenizer with the precision of 0, 1% as well as 3% by weight. Aluminum oxide nanoparticles are produced by means of a catalytic chemical vapor deposition approach. Aluminum oxide nanomaterials are synthesized by utilizing precursors to be mentioned as methane, acetylene, toluene, xylene, benzene, etc. However, it is possible to apply any kind of hydrocarbon or carbonaceous materials for the synthesis of Aluminium oxide nanomaterials.

Figure 2. TEM photograph(Left)and XRD Pattern (Right) of Alminium oxide nano-particles.



Experimental Setup

The schematic diagram of the heat pipe under consideration is demonstrated in Figure 3. The loop is made of a test part, cool water loop as well as data recording system. The water flow at a fixed amount is considered to be 1.7 L/min. The inlet temperature of the water is controlled and also is set at the temperature of $20 \pm 0.5^\circ\text{C}$ by utilizing a thermostatic bath for achieving a steady working temperature during the time that the heat input increases. The close-loop of the evaporator is designed in such a manner that it has the cold water inside the storage tank with 0.125 m^3 volume and water pump. Experiments are done at different heat fluxes as well as three various tilt angles. Position of the heat pipe tilt angle is displayed in Figure 3. Also the temperature of the laboratory is adjust to $25 \pm 1^\circ\text{C}$. The heat pipe which is used in the experiment is a straight copper tube having outer diameter 8 mm as well as length of 190 mm, also has 1 mm wick-thickness copper powder (140-200 nm) sintered wicked. The implemented heat pipe is made of three parts to be mentioned as an evaporator, an adiabatic part as well as a condenser. The condenser part of the plain heat pipe is placed vertically into the cooling chamber for removing the heat from the condenser part with the help of forced convection to a steady temperature bath. In order to heat the evaporator part, a 0-220 V AC electric heater is utilized. Electric power acquires with the measurement of the current imparted to the electric heater as well as the voltage across the heating resistor. In order to reduce the contact thermal resistance among the heater as well as the heat pipe, the electrical heater is glued with thermal grease master cooler-HTX3-GL.

Test Scheme

Once the power supply is turned on, the power is raised. Nearly 15-25 min are needed for the tests in order to attain steady state. The moment that the steady-state condition is attained, the temperature distribution in the heat pipe along with the other experimental parameters is measured and recorded. Henceforth,

The local heat pipe temperature can be evaluated by utilizing four isolated type-T thermocouples. A digital temperature recorder is implemented in such a way that all thermocouples are linked to it. Thermocouples (having the uncertainty less than 0.1 °K) are spread along the surfaces of the heat pipe section in such a way that one thermocouple is connected to the center at evaporator part, two thermocouples are connected to adiabatic part, also one thermocouple is connected to condenser part. Furthermore, a heater Bar (maximum 90W) is utilized as a heat source in the heating part. Therefore, the heating load (Q) as well as the temperature differences (ΔT) are calculated. The thermal resistance associated with the thermosyphon heat pipe is defined as follows,

$$R = \frac{\Delta T}{O} \left[\frac{\circ K}{W} \right]. \quad (1)$$

$$\eta = m_w C_p (T_{wout} - T_{win}) / (V.I). \quad (2)$$

where m_w is taken to be the mass of water, C_p is taken to be the specific heat capacity of water, T_{wout} and T_{win} are the temperature of water in output and input of condenser respectively, also V and I are considered as the voltage and the current of heat source in evaporator part, respectively.

In order to cause the heat pipe to function in the highest capillary pumping pressure, $\Delta P_{c,max}$ should be more than the total pressure fall in the pipe. The pressure fall is constructed of three components which are mentioned as follows,

1. The pressure fall ΔP_l needed to return the liquid from the condenser to the evaporator.
2. The pressure fall ΔP_v needed to make the vapour to flow from the evaporator to the condenser.
3. The pressure because of the gravitational head, ΔP_g , that can be zero, positive or negative, related to the inclination of the heat pipe.

The following relation is extracted (Tien & Rohani, 1974),

$$\Delta p_{c,max} \geq \Delta p_l + \Delta p_v + \Delta p_g. \quad (3)$$

In a case that Equation (3) is not hold, the wick will dry out in the evaporator area also the heat pipe will not function. The highest permitted amount of the heat flux for holding the equation (3) is considered as the capillary limit. Generally, the capillary limit determines the highest amount of heat flux over much of the function domain, nevertheless, the deviser should investigate that a heat pipe is not needed to operate outside the envelope either at design conditions or at start-up.

RESULTS AND DISCUSSIONS

The shape of the heat pipe as well as the working fluid has an important consequence on its efficiency. Hence, the proposed experiment in this book chapter considers different conditions of fluid for evaluating the temperature distribution of heat pipe where the condenser as well as evaporator is perpendicular line pipe. Figure 4 demonstrates the wall temperature distribution based on the length of the heat pipe axial and under the water cooling in a case that Al_2O_3 nanoparticles adds into the working fluid. The increment in the heat pipe wall temperature variance is less than the pure water filled heat pipe under different heat loads.

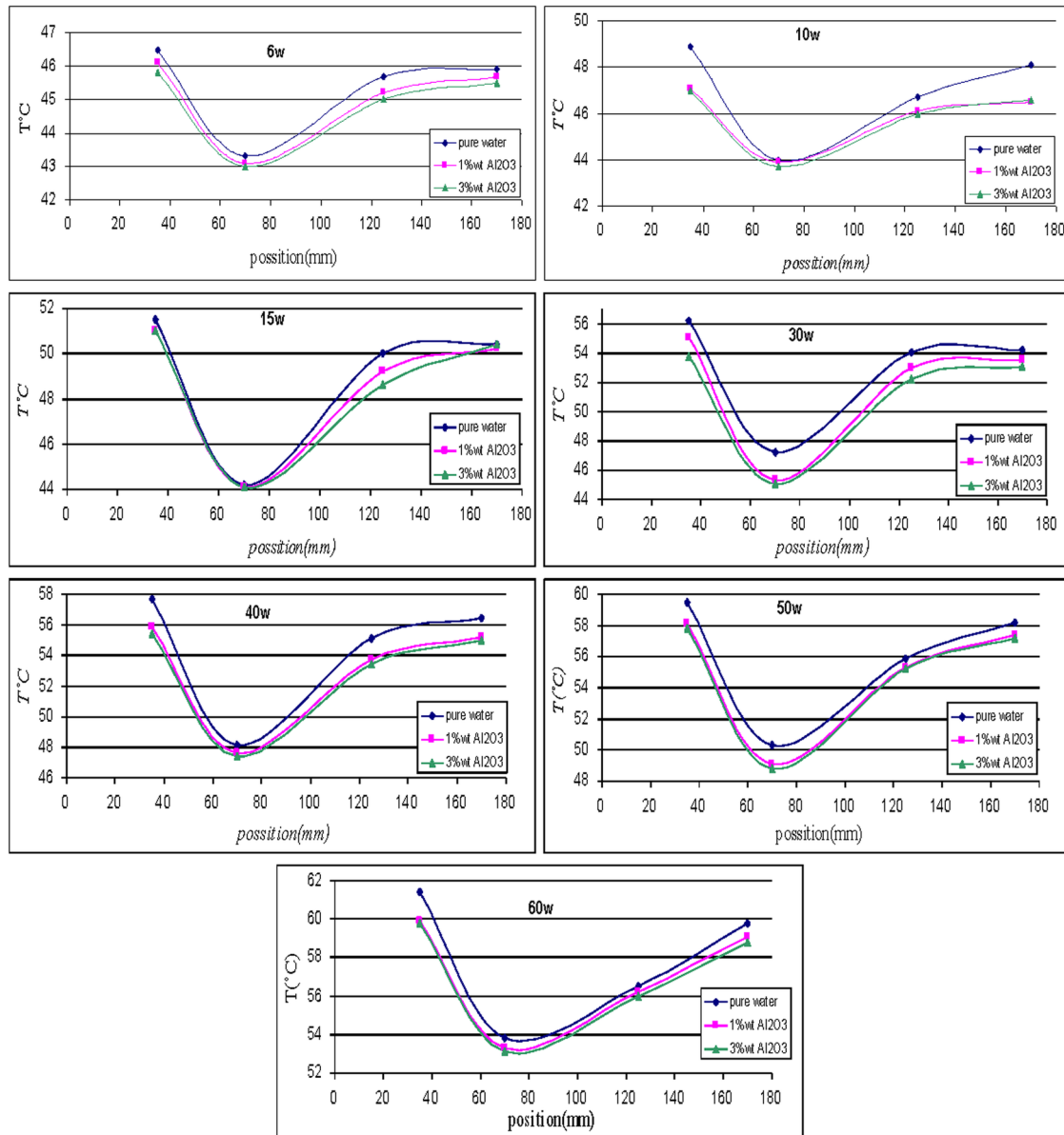
Figure 5. displays the consequence of nanoparticle on the temperature variance of the heat pipe under different input powers. It can be seen from this Figure that by increment in the heat flux, a sudden decrease happens which can cause an improvement in the valve of vapor as well as the rate of transnational speed among condenser and evaporator. Figure 6. displays evaluated value of thermal resistance of the heat pipe with nanofluid as well as the decrement of the thermal resistance by increment in the input power while the nanofluid is added.

The cause of the increment in the heat pipe thermal can be given as below:

Implementing of the existing heat pipe as well as nanofluid theories, especially the ones associated with the boiling limit, can cause a decrement in the boiling limit. The reason of this decrement is the effect of the increment of the liquid conductance in heat pipe. Hence, the thermal resistance of the heat pipe has decrement for the same cause. The high efficiency of the nanofluids demonstrates nanofluid potential as substitutions conventional pure water in grooved heat pipe.

The increment of the convective heat transfer coefficient is associated with the increment of the fluid thermal conductivity. Also, the increment of the thermal conductivity of the nanofluids depends on the increment of the volume concentrations. Adding nanoparticles to liquid considerably makes better the energy transport procedure of the base liquid. A smaller particle size would provide a much larger surface

Figure 4. Average temperature of heat pipe distribution in different heat load.



area for molecular collisions and therefore increase the rate of reaction, making it a better catalyst and reactant. A tiny particle size can generate a great surface region for molecular collisions, also can make an increment in the amount of reaction, causing it a superior catalyst as well as reactant.

The shape of heat pipe, working fluid as well as the consequence of gravity has a significant effect on its performance. This experiment implements various conditions for fluid to evaluate the temperature distribution of heat pipe as well as cooling water evaporator and condenser. In vertical form the vapor bubbles that get heat in the evaporator increase in size. Their own buoyancy make them to grow in the tube part. The growing bubbles in the tube makes the liquid slugs to move downwards unto the evaporator with the help of the gravity force. In the horizontal form (90° inclination angle) of function, there exists no macro motion of bubbles. This can make a conclusion that the gravity has a role in the heat pipe. As

Figure 5. Effect of nanoparticle on the temperature difference of heat pipe under different input power.

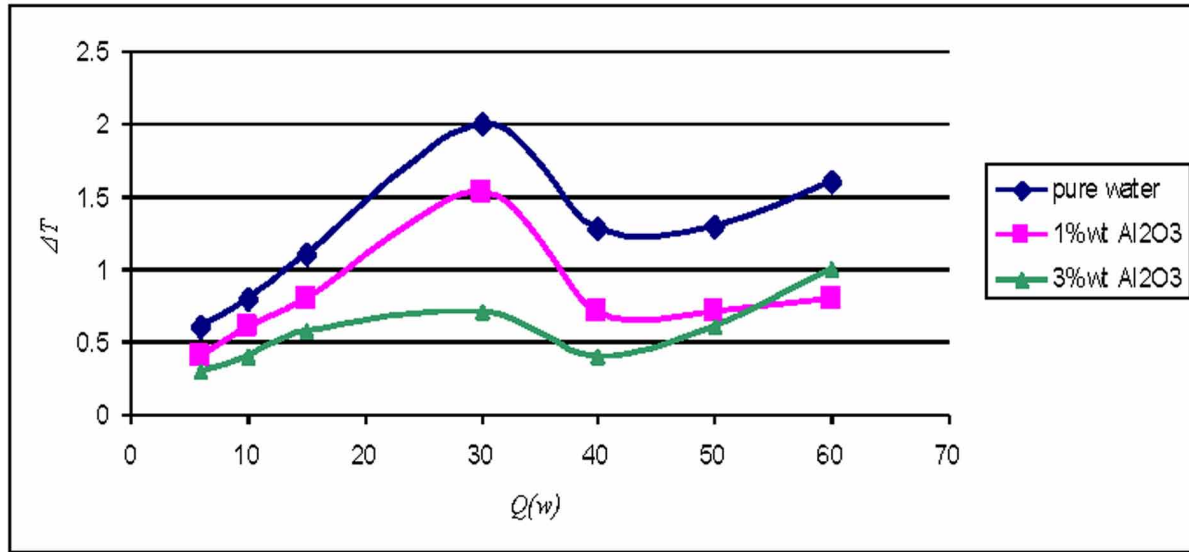
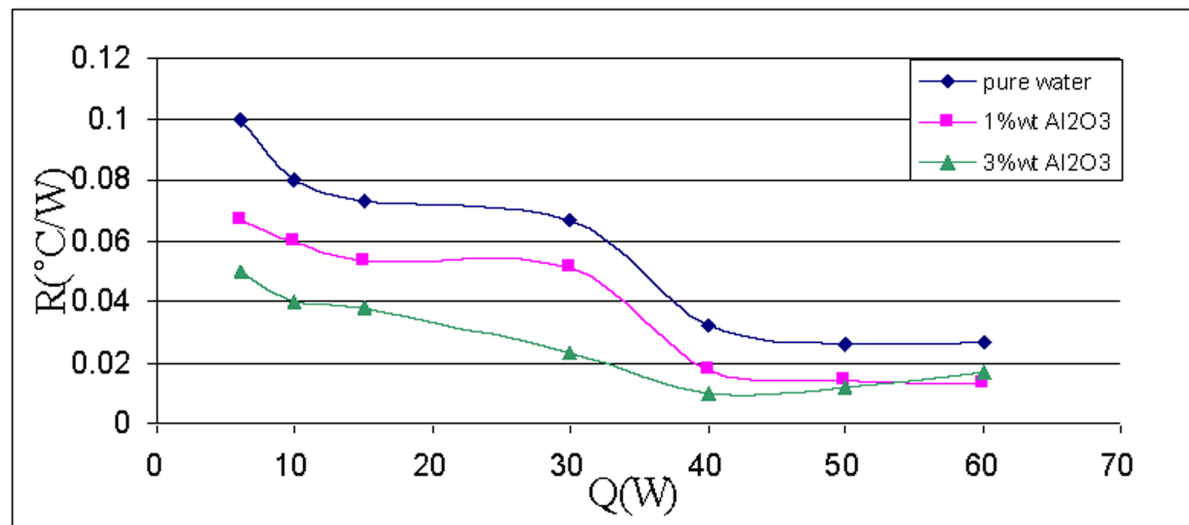


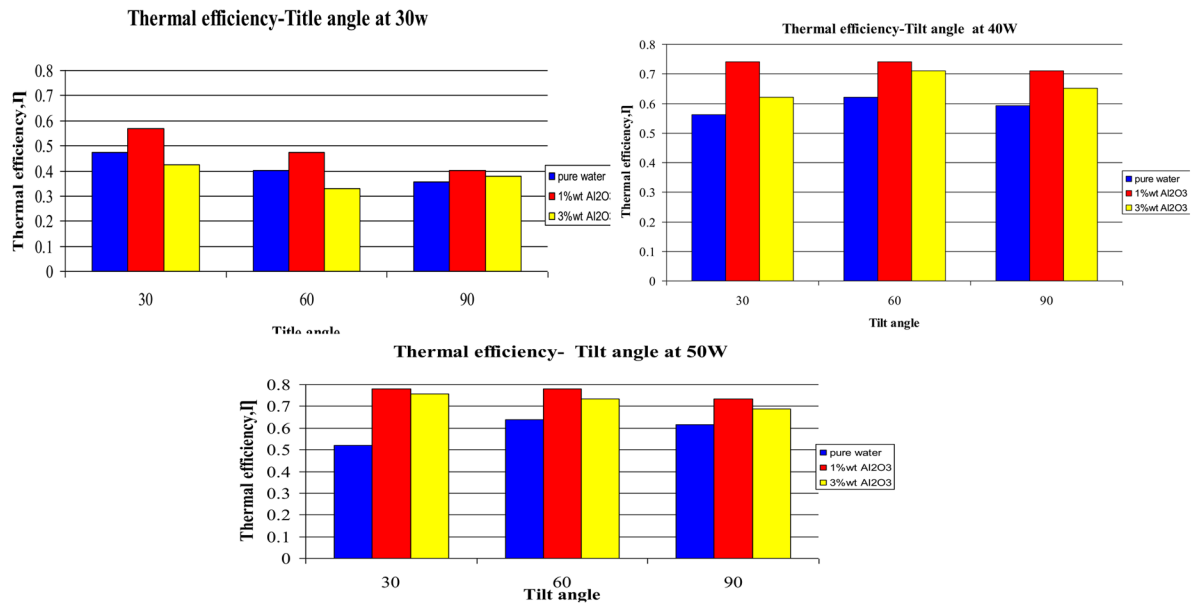
Figure 6. Measured value of thermal resistance of heat pipe with nanofluid



the gravity force has no effect, all the movements related to the bubbles as well as slugs should be carry on with the help of pressure forces. These forces are generated because of the temperature variation which are among evaporator as well as condenser.

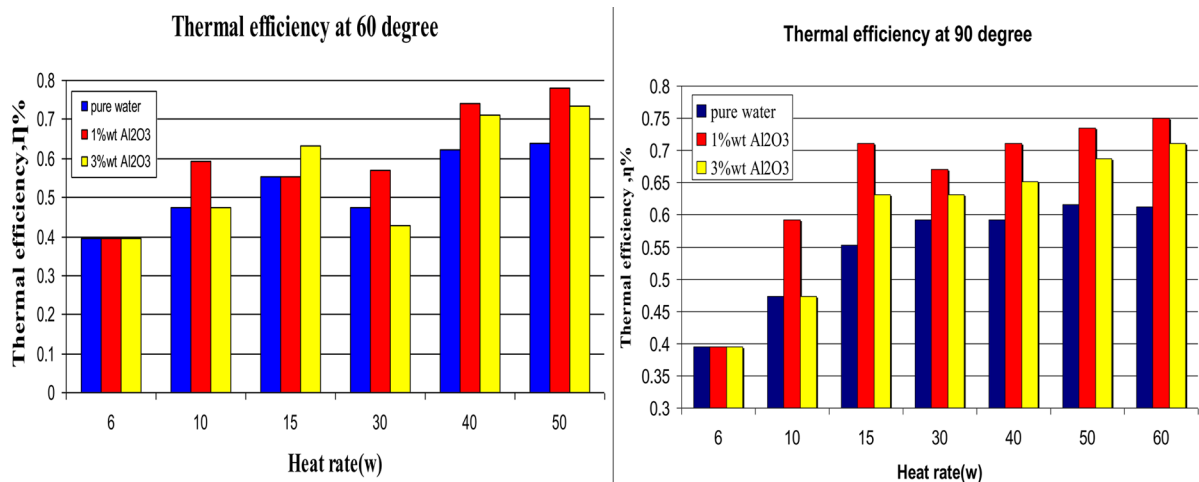
Figure 7 demonstrate the alteration of thermal performance associated with the heat pipe. In order to generate various results, fluids with different concentration percent of Al₂O₃ nanoparticles at 30, 60 as well as 90 tilt angles are utilized. The increasing of the tilt angle can make the increment of flow back into the evaporator part. With increasing of the title angle from 0 to 60 the thermal performance grows. However, in a case that the increasing be more than 60 the thermal performance will reduce. This is due to the impact of incrementing gravitational force among evaporator part as well as condenser

Figure 7. Alteration of thermal performance of heat pipes by different tilt angles and fluids of heat pipes at 30W, 40W and 50W power.



part. In a case that the tilt angle of heat pipe is constant, the performance of the thermal conductivity increases by adding nanoparticles to the base fluid. Adding extra nanoparticles to fluid can transform the property of working fluid at evaporator part into the solid phase. Furthermore, it causes a decrement in the convection efficiency of nanofluid inside evaporator part. Figure 8 demonstrate that at the constant angle of the heat pipe, the increasing of the heat rate will increase the performance of the heat pipe. It is clear from these Figures that after the tilt angle of 60° (Figure 8) the heat pipe performance is reduced. From these Figures it can be seen that the contact angle makes the condensed vapor to return rapidly into the evaporator part with gravity also, less thermal resistances as well as better heat transfer coefficients are extracted.

Figure 8. Variations of Thermal efficiency in different heat rate at 60° and 90° tilt.



FUTURE RESEARCH DIRECTIONS

The investigation of the thermal enhancement of the heat pipe proposed by different nanofluid will be the further work.

CONCLUSION

In this book chapter, an experimental research of the thermal enhancement related to the heat pipe is carried out proficiently by adding Al_2O_3 nano powder to pure water in side 1mm wick-thickness sintered circular heat pipe in various title angle.

Important conclusions have been obtained and are summarized as following:

- The ideal value of thermal efficiency obtains at the concentration of 1.0% wt nanofluid.
- The heat transfer coefficient increases with increasing the applied heat flux at the evaporator.
- Adding Al_2O_3 nanoparticles to pure water could enhance the thermal efficiency of heat pipe.
- With enhancement of the heat flux, the heat pipe efficiency tends to increase.
- The lowest amount of temperature in the heat pipe can be observed in the curved zone and this amount of temperature is decreased from the evaporator to the curve and then increase.
- Effects of heat pipe tilt angle on the thermal efficiency are checked and is shown that it can be increased at higher tilt angle.
- The inclination angle has a great effect on the heat pipe thermal resistance using water as the working fluid.
- The heat pipe thermal efficiency get enhanced with increasing nanoparticles concentrations and tilt angles.

REFERENCES

- Alizad, K., Vafai, K., & Shafahi, M. (2012). Thermal performance and operational attributes of the startup characteristics of flat shaped heat pipes using nanofluids. *International Journal of Heat and Mass Transfer*, 55(1-3), 140–155. doi:10.1016/j.ijheatmasstransfer.2011.08.050
- Chandrasekar, M., Suresh, S., & Chandra Bose, A. (2010). Experimental investigations and theoretical determination of thermal conductivity and viscosity of Al_2O_3 /water nanofluid. *Experimental Thermal and Fluid Science*, 34(2), 210–216. doi:10.1016/j.expthermflusci.2009.10.022
- Chen, Y., Wang, P., & Liu, Z. (2013b). Application of water-based SiO_2 functionalized nanofluid in a loop thermosyphon. *International Journal of Heat and Mass Transfer*, 56(1-2), 59–68. doi:10.1016/j.ijheatmasstransfer.2012.09.048
- Chen, Y. J., Wang, P. Y., Liu, Z. H., & Li, Y. Y. (2013a). Heat transfer characteristics of a new type of copper wire-bonded flat heat pipe using nanofluids. *International Journal of Heat and Mass Transfer*, 67, 548–559. doi:10.1016/j.ijheatmasstransfer.2013.08.060
- Choi, S. (1995). Enhancing thermal conductivity of fluids with nano-particles. ASME, 231.

- Choi, S. (1998). Nanofluid technology: Current status and future research. *Energy*, 26.
- Choi, S. U. S., Wang, X., & Xu, W. (1999). Thermal conductivity of nano-particle-fluidmixture. *Journal of Thermophysics and Heat Transfer*, 13(4), 474–480. doi:10.2514/2.6486
- Do, K. H., Ha, H. J., & Jang, S. P. (2010b). Thermal resistance of screen mesh wick heat pipes using the water-based Al₂O₃ nanofluids. *International Communications in Heat and Mass Transfer*, 53(25-26), 5888–5894. doi:10.1016/j.ijheatmasstransfer.2010.07.050
- Do, K. H., Jang, S. P., Hyung, K., Pil, S., Do, K. H., & Jang, S. P. (2010a). Effect of nanofluids on the thermal performance of a flat micro heat pipe with a rectangular grooved wick. *International Journal of Heat and Mass Transfer*, 53(9-10), 2183–2192. doi:10.1016/j.ijheatmasstransfer.2009.12.020
- Eastman, J. A., Choi, S. U. S., Li, S., Yu, W., & Thompson, L. J. (2001). Anomalously increased effective thermal conductivities of ethylene glycol-based nanofluids containing copper nanoparticles. *Applied Physics Letters*, 78(6), 718–720. doi:10.1063/1.1341218
- Gaugler, R. S. (1944). Heat transfer devices. *U.S. patent*, 2, 350-348.
- Ghanbarpour, M., & Khodabandeh, R. (2015a). Entropy generation analysis of cylindrical heat pipe using nanofluid. *Thermochimica Acta*, 610, 37–46. doi:10.1016/j.tca.2015.04.028
- Ghanbarpour, M., Nikkam, N., Khodabandeh, R., Toprak, M. S., & Muhammed, M. (2015b). Thermal performance of screen mesh heat pipe with Al₂O₃ nanofluid. *Experimental Thermal and Fluid Science*, 66, 213–220. doi:10.1016/j.expthermflusci.2015.03.024
- Goshayeshi, H. R., & Chaer, I. (2016). Experimental study and flow visualization of Fe₂O₃/Kerosene in glass oscillating heat pipes. *Applied Thermal Engineering*, 103, 1213–1218. doi:10.1016/j.applthermaleng.2016.04.109
- Goshayeshi, H. R., Goodarzi, M., & Dahari, M. (2015). Effect of magnetic field on the heat transfer rate of kerosene / Fe₂O₃ nanofluid in a copper oscillating heat pipe. *Experimental Thermal and Fluid Science*, 68, 663–668. doi:10.1016/j.expthermflusci.2015.07.014
- Grover, G. M., Cotter, T. P., & Erikson, G. F. (1964). Structures of very high thermal conductivity. *Journal of Applied Physics*.
- Gunnasegaran, P., Abdullah, M. Z., & Shuaib, N. H. (2013). Influence of nanofluid on heat transfer in a loop heat pipe. *International Communications in Heat and Mass Transfer*, 47, 82–91. doi:10.1016/j.icheatmasstransfer.2013.07.003
- Gunnasegaran, P., Zulkifly, M., & Zamri, M. (2015). Optimization of SiO₂ nanoparticle mass concentration and heat input on a loop heat pipe. *Case Studies in Thermal Engineering*, 6, 238–250. doi:10.1016/j.csite.2015.10.004
- Heris, S. Z., Mohammadpur, F., & Shakouri, A. (2014). Effect of electric field on thermal performance of thermosyphon heat pipes using nanofluids. *Materials Research Bulletin*, 53, 21–27. doi:10.1016/j.materresbull.2014.01.030
- Huminic, G., & Huminic, A. (2013). Numerical study on heat transfer characteristics of thermosyphon heat pipes using nanofluids. *Energy Conversion and Management*, 76, 393–399. doi:10.1016/j.enconman.2013.07.026

Hung, Y. H., Teng, T. P., & Lin, B. G. (2012). Evaluation of the thermal performance of a heat pipe using alumina nanofluids. *Experimental Thermal and Fluid Science*, 44, 504–511. doi:10.1016/j.exp-thermflusci.2012.08.012

Jafari, R., & Yu. W. (2015). Artificial neural network approach for solving strongly degenerate parabolic and burgers-fisher equations. *12th International Conference on Electrical Engineering, Computing Science and Automatic Control*. 10.1109/ICEEE.2015.7357914

Jafari, R., & Yu. W. (2017). Uncertain nonlinear system control with fuzzy differential equations and Z-numbers. *18th IEEE International Conference on Industrial Technology*, 1, 890-895. 10.1109/ICIT.2017.7915477

Jafari, R., & Razvarz, S. (2017). Solution of Fuzzy Differential Equations using Fuzzy Sumudu Transforms. *IEEE International Conference on Innovations in Intelligent Systems and Applications*, 84-89. 10.1109/INISTA.2017.8001137

Jafari, R., & Razvarz, S. (2018). Solution of fuzzy differential equations using fuzzy sumudu transforms. *Mathematical and Computational Applications*, 1-15.

Jafari, R., Razvarz, S., & Gegov, A. (2018a). A New Computational Method for Solving Fully Fuzzy Non-linear Systems. *Lecture Notes in Computer Science*, 11055, 503-512. doi:10.1007/978-3-319-98443-8_46

Jafari, R., Razvarz, S., & Gegov, A. (2019a). Neural Network Approach to Solving Fuzzy Nonlinear Equations using Z-Numbers. *IEEE Transactions on Fuzzy Systems*, 1. Advance online publication. doi:10.1109/TFUZZ.2019.2940919

Jafari, R., Razvarz, S., & Gegov, A. (2019b). Solving Differential Equations with Z-Numbers by Utilizing Fuzzy Sumudu Transform. *Intelligent Systems and Applications*. IntelliSys 2018. Advances in Intelligent Systems and Computing, 869, 1125-1138.

Jafari, R., Razvarz, S., Gegov, A., & Paul, S. (2018b). Fuzzy modeling for uncertain nonlinear systems using fuzzy equations and Z-numbers. *Advances in Computational Intelligence Systems: Contributions Advances in Intelligent Systems and Computing*, 840, 66-107.

Jafari, R. Yu. W., & Li, X. (2016). Solving Fuzzy Differential Equation with Bernstein Neural Networks. *IEEE International Conference on Systems, Man, and Cybernetics*, 1245-1250. 10.1109/SMC.2016.7844412

Jafarian, A., Measoomynia, S., & Jafari, R. (2012). Solving Fuzzy Equations Using Neural Nets with a New Learning Algorithm. *Journal of Advances in Computer Research*, 3, 33–45.

Ji, Y., Liu, G., Ma, H., Li, G., & Sun, Y. (2013). An experimental investigation of heat transfer performance in a polydimethylsiloxane (PDMS) oscillating heat pipe. *Applied Thermal Engineering*, 61(2), 690–697. doi:10.1016/j.applthermaleng.2013.09.001

Kamyar, A. K., Ong, S., & Saidur, R. (2013). Effects of nanofluids on heat transfer characteristics of a two-phase closed thermosyphon. *International Journal of Heat and Mass Transfer*, 65, 610–618. doi:10.1016/j.ijheatmasstransfer.2013.06.046

Kebllinski, P., Eastman, J. A., & Cahill, D. G. (2005). Nanofluids for thermal transport. *Materials Today*, 8(6), 36–44. doi:10.1016/S1369-7021(05)70936-6

- Kole, M., & Dey, T. K. (2013). Thermal performance of screen mesh wick heat pipes using water-based copper nanofluids. *Applied Thermal Engineering*, 20(1), 763–770. doi:10.1016/j.applthermaleng.2012.06.049
- Lin, Y., Kang, S., & Chen, H. (2008). Effect of silver nanofluid on pulsating heat pipe thermal performance. *Applied Thermal Engineering*, 28(11–12), 1312–1317. doi:10.1016/j.applthermaleng.2007.10.019
- Liu, Z., Li, Y., & Bao, R. (2010). Thermal performance of inclined grooved heat pipes using nanofluids. *International Journal of Thermal Sciences*, 49(9), 1680–1687. doi:10.1016/j.ijthermalsci.2010.03.006
- Liu, Z., Li, Y., & Bao, R. (2011b). Compositive effect of nanoparticle parameter on thermal performance of cylindrical microgrooved heat pipe using nanofluids. *International Journal of Thermal Sciences*, 50(4), 558–568. doi:10.1016/j.ijthermalsci.2010.11.013
- Liu, Z., Xiong, J., & Bao, R. (2007). Boiling heat transfer characteristics of nanofluids in a flat heat pipe evaporator with micro grooved surface. *International Journal of Multiphase Flow*, 33(12), 1284–1295. doi:10.1016/j.ijmultiphaseflow.2007.06.009
- Liu, Z., & Zhu, Q. (2011a). Application of aqueous nanofluids in a horizontal mesh heat pipe. *Energy Conversion and Management*, 52(1), 292–300. doi:10.1016/j.enconman.2010.07.001
- Liu, Z. H., Hu, R. L., Lu, L., Zhao, F., & Xiao, H. (2013). Thermal performance of an open thermosyphon using nanofluid for evacuated tubular high temperature air solar collector. *Energy Conversion and Management*, 73, 135–143. doi:10.1016/j.enconman.2013.04.010
- Mashaei, P. R., Shahryari, M., Fazeli, H., & Hosseinalipour, S. M. (2016). Numerical simulation of nanofluid application in a horizontal mesh heat pipe with multiple heat sources: A smart fluid for high efficiency thermal system. *Applied Thermal Engineering*, 100, 1016–1030. doi:10.1016/j.applthermaleng.2016.02.111
- Menlik, T., Sözen, A., & Gürü Öztaş, M. S. (2014). Heat transfer enhancement using MgO/water nanofluid in heat pipe. *Journal of the Energy Institute*. doi:10.1016/j.joei.2014.10.001
- Murshed, S. M. S., Leong, K. C., & Yang, C. (2005). Enhanced thermal conductivity of TiO₂ - Water based nanofluids. *International Journal of Thermal Sciences*, 44(4), 367–373. doi:10.1016/j.ijthermalsci.2004.12.005
- Naveen Kumar Gupta, A. B. (2019). Numerical study of CeO₂/H₂O nanofluid application on thermal performance of heat pipe. *Materials Today: Proceedings*, 18, 1006–1016. doi:10.1016/j.matpr.2019.06.541
- Noie, S. H., Heris, S. Z., Kahani, M., & Nowee, S. M. (2009). Heat transfer enhancement using Al₂O₃/water nanofluid in a two-phase closed thermosyphon. *International Journal of Heat and Fluid Flow*, 30(4), 700–705. doi:10.1016/j.ijheatfluidflow.2009.03.001
- Pinto, R. V., & Fiorelli, F. A. (2019). Analysis of behaviour of computational model to evaluate performance of heat pipe containing nanofluids. *Journal of Central South University*, 26(5), 1306–1326. doi:10.1007/11771-019-4089-4
- Putra, N., Saleh, R., Nata, W., Okta, A., & Hamid, Z. (2014). Thermal performance of biomaterial wick loop heat pipes with water-base Al₂O₃ nanofluids. *International Journal of Thermal Sciences*, 76, 128–136. doi:10.1016/j.ijthermalsci.2013.08.020

Putra, N., Septiadi, W. N., Rahman, H., & Irwansyah, R. (2012). Thermal performance of screen mesh wick heat pipes with nanofluids. *Experimental Thermal and Fluid Science*, 40, 10–17. doi:10.1016/j.expthermflusci.2012.01.007

Qu, J., & Wu, H. (2011). Thermal performance comparison of oscillating heat pipes with SiO_2 / water and Al_2O_3 / water nanofluids. *International Journal of Thermal Sciences*, 50(10), 1954–1962. doi:10.1016/j.ijthermalsci.2011.04.004

Qu, J., Wu, H., & Cheng, P. (2010). Thermal performance of an oscillating heat pipe with Al_2O_3 –water nanofluids. *International Communications in Heat and Mass Transfer*, 37(2), 111–115. doi:10.1016/j.icheatmasstransfer.2009.10.001

Rakesh, K., & Bumataria, N. C. (2019). Current research aspects in mono and hybrid nanofluid based heat pipe technologies. *Heliyon*. PMID:31193568

Razvarz, Vargas-Jarillo, & Jafari. (2019a). Pipeline Monitoring Architecture Based on Observability and Controllability Analysis. *IEEE International Conference on Mechatronics (ICM)*, 1, 420–423. 10.1109/ICMECH.2019.8722875

Razvarz, S., & Jafari, R. (2016). Experimental Study of Al_2O_3 Nanofluids on the Thermal Efficiency of Curved Heat Pipe at Different Tilt Angle. *2nd international congress on technology engineering and science (ICONTEs)*.

Razvarz, S., & Jafari, R. (2017a). ICA and ANN Modeling for Photocatalytic Removal of Pollution in Wastewater. *Mathematical and Computational Applications*, 22(3), 38–48. doi:10.3390/mca22030038

Razvarz, S., & Jafari, R. (2017b). Intelligent Techniques for Photocatalytic Removal of Pollution in Wastewater. *Journal of Electrical Engineering*, 5, 321–328. doi:10.17265/2328- 2223/2017.06.004

Razvarz, S., & Jafari, R. (2018). Experimental study of Al_2O_3 nanofluids on the thermal efficiency of curved heat pipe at different tilt angle. *Journal of Nanomaterials*, 2018, 1–7. doi:10.1155/2018/1591247

Razvarz, S., Jafari, R., Gegov, A., Yu, W., & Paul, S. (2018). Neural network approach to solving fully fuzzy nonlinear systems. In *Fuzzy modeling and control Methods Application and Research*. Nova Science Publisher, Inc.

Razvarz, S., Jafari, R., Yu, W., & Khalili, A. (2017). PSO and NN Modeling for Photocatalytic Removal of Pollution in Wastewater. *2017 14th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 20, 1–6.

Razvarz, S., Vargas-jarillo, C., Jafari, R., & Gegov, A. (2019b). Flow Control of Fluid in Pipelines Using PID Controller. *IEEE Access: Practical Innovations, Open Solutions*, 7, 25673–25680. doi:10.1109/ACCESS.2019.2897992

Ren, Q. (2019). Enhancement of nanoparticle-phase change material melting performance using a sinusoidal heat pipe. *Energy Conversion and Management*, 180, 784–795. doi:10.1016/j.enconman.2018.11.033

Riehl, R. R., & Santos, N. (2012). Water-copper nanofluid application in an open loop pulsating heat pipe. *Applied Thermal Engineering*, 42, 6–10. doi:10.1016/j.applthermaleng.2011.01.017

Saidur, R., Leong, K. Y., & Mohammad, H. A. (2011). A review of applications and challenges of nanofluids. *Renewable & Sustainable Energy Reviews*, 15(3), 1646–1668. doi:10.1016/j.rser.2010.11.035

- Saleh, R., Putra, N., Purbo, S., & Nata, W. (2013). Experimental investigation of thermal conductivity and heat pipe thermal performance of ZnO nanofluids. *International Journal of Thermal Sciences*, 63, 125–132. doi:10.1016/j.ijthermalsci.2012.07.011
- Sekhon, B. (2014). Nanotechnology in agri-food production: An overview. *Nanotechnology, Science and Applications*, 7, 31–35. doi:10.2147/NSA.S39406 PMID:24966671
- Senthil, R., Ratchagaraja, D., Silambarasan, R., & Manikandan, R. (2016). Contemplation of thermal characteristics by filling ratio of Al_2O_3 nanofluid in wire mesh heat pipe. *Alexandria Engineering Journal*, 30(2), 2–7. doi:10.1016/j.aej.2016.03.011
- Senthilkumar, R., Vaidyanathan, S., & Sivaraman, B. (2012). Effect of Inclination Angle in Heat Pipe Performance Using Copper Nanofluid. *Procedia Engineering*, 38, 3715–3721. doi:10.1016/j.pro-eng.2012.06.427
- Shafahi, M., Bianco, V., Vafai, K., & Manca, O. (2010). Thermal performance of flat-shaped heat pipes using nanofluids. *International Journal of Heat and Mass Transfer*, 53(7-8), 1438–1445. doi:10.1016/j.ijheatmasstransfer.2009.12.007
- Singh, A. K., & Raykar, V. S. (2008). Microwave synthesis of silver nanofluids with poly vinyl pyrrolidone (PVP) and their transport properties. *Colloid & Polymer Science*, 286(14-15), 1667–1673. doi:10.1007/00396-008-1932-9
- Solomon, A. B., Ramachandran, K., Asirvatham, L. G., & Pillai, B. C. (2014). Numerical analysis of a screen mesh wick heat pipe with Cu / water nanofluid. *International Journal of Heat and Mass Transfer*, 75, 523–533. doi:10.1016/j.ijheatmasstransfer.2014.04.007
- Solomon, A. B., Ramachandran, K., & Pillai, B. C. (2012). Thermal performance of a heat pipe with nanoparticles coated wick. *Applied Thermal Engineering*, 36, 106–112. doi:10.1016/j.applthermaleng.2011.12.004
- Sözen, A., Menlik, T., Gürü, M., Boran, K., Kılıç, F., Aktaş, M., & Çakır, M. T. (2016). A comparative investigation on the effect of flyash and alumina nanofluids on the thermal performance of two-phase closed thermo-syphon heat pipes. *Applied Thermal Engineering*, 96, 330–337. doi:10.1016/j.applthermaleng.2015.11.038
- Tien, C. L., & Rohani, A. R. (1974). Analysis of the effects of vapor pressure drop on heat pipe performance. *International Journal of Heat and Mass Transfer*, 17(1), 61–67. doi:10.1016/0017-9310(74)90038-6
- Trefethen, L. (1962). On the surface tension pumping of liquids or a possible role of the candlewick in space exploration. GE Tech. Int. Ser. No G15-D114, General Electric Co.
- Venkatachalapathy, S., Kumaresan, G., & Suresh, S. (2015). Performance analysis of cylindrical heat pipe using nanofluids – An experimental study. *International Journal of Multiphase Flow*, 72, 188–197. doi:10.1016/j.ijmultiphaseflow.2015.02.006
- Wan, Z., Deng, J., Li, B., Xu, Y., Wang, X., & Tang, Y. (2015). Thermal performance of a miniature loop heat pipe using water -copper nanofluid. *Applied Thermal Engineering*, 78, 712–719. doi:10.1016/j.applthermaleng.2014.11.010

Wang, B., Shih, T. M., Tian, B., Wu, C. X., & Chang, R. R. G. (2017). Mildly zigzag heat transfer affected by aspect ratios for recirculating flows in rectangular enclosures. *International Journal of Heat and Mass Transfer*, 107, 372–378. doi:10.1016/j.ijheatmasstransfer.2016.11.056

Wang, G.-S., Song, B., & Liu, Z. H. (2010). Operation characteristics of cylindrical miniature grooved heat pipe using aqueous CuO nanofluids. *Experimental Thermal and Fluid Science*, 34(8), 1415–1421. doi:10.1016/j.expthermflusci.2010.07.004

Yu, W., & Jafari, R. (2019). *Fuzzy modeling and control with fuzzy equations and z- number*, IEEE Press series on systems science and engineering. Wiley-IEEE Press. John Wiley & Sons, Inc.

KEY TERMS AND DEFINITIONS

Aluminum Oxide: A chemical compound of aluminum and oxygen with the chemical formula Al_2O_3 .

Condenser: A device that cools gases into liquids.

Evaporator: A vessel or stage which turns a liquid into vapor.

Heat Pipe: Tubular device that is very efficient in transferring heat.

Heat Transfer: The movement of thermal energy from one thing to another thing of different temperature.

Nano Materials: Materials of which a single unit is sized between 1 to 1000 nanometers (10^{-9} meter)

Nanofluids: A fluid containing nanometer-sized particles, called nanoparticles.

Thermal Resistance: The ability of a material to resist flow of heat.

Thermosyphon: A device that transfers heat via natural convection in a fluid.

Modeling of the Aerodisperse Systems Hydrodynamics in Devices With Directional Motion of the Fluidized Bed

Artem Artyukhov

Sumy State University, Ukraine

Jan Krmela

 <https://orcid.org/0000-0001-9767-9870>

Alexander Dubcek University of Trencin, Slovakia

Nadiia Artyukhova

Sumy State University, Ukraine

Ruslan Ostroha

Sumy State University, Ukraine

INTRODUCTION

Among various dispersed flows, fluidized bed plays the most significant role in modern technology. The fluidization technique has become widely used due to high intensity of processes.

Disadvantages of the fluidized bed device can include (Philippsen et al., 2015), (Haron et al., 2017), (Huili et al., 2017): different residence time of a particle in the device, a need for thorough cleaning of the exhaust air and material entrainment (in particular it is applicable to small particles of the system). In addition, a significant drawback is the return of fine particles back in the fluidized bed zone. The analysis of various fluidized bed granulation equipment in chemical (Caiyuan et al., 2004), food (Patel et al., 2011) and pharmaceuticals (Agrawal and Naveen, 2011) industries showed the urgent need to organize the mutual flow motion, which will enhance the quality of the final product.

It is rational to use the same device for multiple processes in low-tonnage and multi-assortment production in order to reduce the range of the equipment. Implementation of the new forms to organize mutual flows motion (while keeping the same principles of the fluidized phases contact), which would intensify processing of the dispersed materials without considerable increase in the energy costs, is a promising direction for the development of heat and mass exchange processes in the heterogeneous systems (Artyukhov et al., 2019), (Van Ommen et al., 2012).

Among the major techniques of controlling the polydisperse particles residence time in the device, one should notice the following:

1. To develop a direct movement of particles using accelerating elements (gas distributors of vortex type) (Sklabinskyi and Artyukhov, 2013).
2. To design the device with a variable cross-sectional area applied to the granulation, cooling and dedusting process (Ostroha et al., 2019).

DOI: 10.4018/978-1-7998-3479-3.ch088

3. The promising direction to reduce financial and energy costs on the heat and mass transfer processes in the fluidized bed is the application of sectioning (vertical and horizontal) to create different conditions for the particle heightwise (lengthwise) motion in the device (Yukhymenko et al., 2016).

BACKGROUND

There are various possibilities to apply suspended layer (fluidized bed, weighted layer). There are many successful cases regarding the introduction of this method in industrial practice, while in others it is at the stage of laboratory research. The processes taking place in heterogeneous systems using the suspended layer method have a great industrial application. It can be used to stove sulphide, arsenic and antimony ores to facilitate the extraction of gold or silver, for pyrite and pyrrhotine roasting to obtain SO_2 in the production of sulfuric acid.

The suspended layer is in great demand in metallurgy for stoving of copper, cobalt and zinc sulfide ores to obtain valuable metals.

Fluidized bed devices for drying solid materials (coal, cement, limestone, etc.) are known in the whole world. Economic considerations make the use of these devices particularly interesting when large-tonnage materials are to be processed. A suspended layer dryer can also be used as a classifier since the drying and classification processes take places simultaneously in the device.

The way to obtain granulation products in the fluidized bed (suspended layer) is used by the world well-known manufacturers of fertilizers and pharmaceutical products: Urea Casale S. A. (Switzerland), Kahl Group (Germany), Stamicarbon (Netherlands), Toyo Engineering Corporation (Japan), Changzhou Xianfeng Drying Equipment Company Ltd (China) Glatt (Germany), Uhde Fertilizer Technology (Netherlands), Rottendorf Pharma (Germany) (Saikh et al, 2013).

The world practice of recent years in the small power engineering has shown that the fluidized bed technology is effective for the utilization of different biomass types, industrial waste and effluents.

The development and implementation of fluidized bed devices are held back in many industries by the lack of reliable methods for calculating them. One of the main reasons is the lack of knowledge of complex and varied processes in the fluidized bed.

When constructing a mathematical model, one should take into account that there is a huge amount of particles in the operating device (Palappan and Sai, 2008). These particles are of different sizes and they move at a different rate in all directions (Ostroha et al., 2017).

Mathematical modeling lets to establish and to optimize mode and technological parameters of the investigated process, as well as to minimize the use of additional automation tools (Trojan, 2015), (Zimmermann and Taghipour, 2015), (Zhaofeng, 2006).

The flows motion hydrodynamics modeling is carried out by the finite-volume method (a numerical method of the differential equations system integration in the differential derivatives). An important benefit of the finite-volume method is that the conservation law of integral values (of the flow rate, of motion quantity) is implemented on every chamber of the computational grid, but not only within it, due to the strong thickening of the computational grid. A closed area of the fluid or gas flow is selected in the calculation. The macroscopic values fields (in this paper – velocity), describing the environment in time and satisfying the definite laws, formed mathematically, are searched for it. The conservation laws in Eulerian variables are mostly used. This model is the most common and the most complicated among multi-phase flow models. The substance of every phase is a solid phase, and the substance motion of

every phase is modeled with Navier-Stokes (Reynolds) equations system, equation of continuity and energy. According to this model, the motion equations for every phase, are solved together.

In order to carry out modeling, it is convenient to solve the above equations by the numerical methods (DEM - Discrete Element Method), identifying a discrete set of values in the proper place (mesh, node of grid), space (during the stationary regimes of the flows motion) instead of the continuous solution. In order to achieve the maximum accuracy of solutions, that method to demonstrate discrete values is chosen, which due to the discretization complies with algebraic equations analogues. As a result, the mathematical problem to solve the differential or integral equations system can be reduced to the algebraic equations system.

There is also a method to describe the axisymmetric motion of the swirling gas flow, based on a system of Navier-Stokes equations (1) and the continuity equation (2) (Artyukhov and Sklabinskyi, 2015):

$$\frac{\partial}{\partial t}(\rho_g V_i) + \frac{\partial}{\partial q_j}(\rho_g V_i V_j) = -\frac{\partial p}{\partial q_i} + \frac{\partial}{\partial q_j} \left[\varepsilon \left(\frac{\partial V_i}{\partial q_j} + \frac{\partial V_j}{\partial q_i} \right) \right] + f_i, \quad (1)$$

$$\frac{\partial \rho_g}{\partial t} + \frac{\partial}{\partial q_j}(\rho V_j) = 0, \quad (2)$$

where ε is a coefficient of turbulent viscosity; t – time; V_i – instantaneous component of the velocity on the given axis; ρ_g – gas density; p – pressure; f_i – component, characterizing the action of the bulk forces; q_j – axes of coordinates.

Analysis of the previous works in the two-phase flows modeling sector, which consists of gas as a disperse phase and disperse particles, showed that one of the most perspective methods to calculate the motion of the particles is a trajectory method (Khanali et al., 2013). Authors (Goldschmidt et al., 2003) concluded that in case of the compressed motion modeling of particles with large diameter (0.5-5 mm), the Lagrangian model regarding the force analysis of the particle motion by the differential equations of motion forms the base for it. The trajectory method with highly accurate results can be used only if there is any software which enables to export theoretical model of the single-particle motion and to consider the flow constraint degree.

The motion of particles in the workspace was simulated on the basis of the differential equations system of the particle motion (Artyukhov et al., 2017):

$$\left. \begin{aligned} m \frac{d^2 q_1}{dt^2} &= -g + \psi \cdot \frac{\pi \cdot \mu_g \cdot d}{8 \cdot m} (V_1 - W_1), \\ m \frac{d^2 q_2}{dt^2} &= \frac{W_3^2}{r} + \psi \cdot \frac{\pi \cdot \mu_g \cdot d}{8 \cdot m} (V_2 - W_2), \\ m \frac{d^2 q_3}{dt^2} &= -\frac{W_2 W_3}{r} + \psi \cdot \frac{\pi \cdot \mu_g \cdot d}{8 \cdot m} (V_3 - W_3) \end{aligned} \right\}, \quad (3)$$

where m – particle's mass; r – a current radius of the workspace in a unit; W_1, W_2, W_3 – radial, circumferential and flow rate components of the particle's velocity correspondingly; g – gravitational acceleration; ψ – a linear coefficient of particle resistance to the gas flow; μ_g – gas flow viscosity; d – a particle's diameter.

Criteria to choose the two-phase system flow model are the following.

1. *Ratio β of the disperse phase substance mass (d) to the mass of the carrier phase substance (c):*

$$\beta = \gamma \frac{F_d}{F_c}, \quad (4)$$

where F_d and F_c – volume fractions, γ – the ratio of the disperse phase density to the carrier phase density, $\gamma = \rho_d / \rho_c$.

If β has high values, the dispersed particles have a great impact on the carrier phase flow, so it is necessary to use only the multi-phase Eulerian model for the proper flow modeling.

2. *Stokes number St :*

$$\beta = \gamma \frac{F_d}{F_c}, \quad (5)$$

where t_d – time, which describes the particles motion, $t_d = (\rho_d d_d^2) / (18 \mu_c)$, d_d – diameter of the particle, μ_c – viscosity of the carrier phase substance, $t_c = L_c / U_c$ – time, which describes the carrier phase, L_c – characteristic length, U_c – characteristic velocity.

The authors (Gidaspow, 1994), (Crowe, 2006), (Sinaiski, 2010), (Marchisio et al., 2013) suggest taking into account the constrained mode of motion when we calculate the particle residence time in the workspace of the device

$$\tau_{AB} = \tau f_{\tau}(\psi), \quad (6)$$

where τ_{cm} – velocity of the constrained motion of particles in the disperse phase; τ – a velocity of a single particle motion; $f_{\tau}(\psi)$ – an empirical influence function of the constraint conditions on the residence time of a particle in the workspace.

The function $f_{\tau}(\psi)$ is:

$$f_{\tau}(\psi) = (1 - \psi)^{-m}, \quad (7)$$

where m – an empirical measure of the stage (a constraint coefficient for the time calculations).

Ansys CFX and Ansys Fluent (<https://www.ansys.com>) software is based on a finite-volume method of solving hydrodynamic equations and using a rectangular adaptive mesh with local grinding. In order to approximate curvilinear geometry with higher accuracy, the technology of geometry grid density is used. This technology enables to import geometry from CAD systems and to share information with finite

element analysis systems. This technology has solved the problem of automatic mesh generation – in order to generate a mesh, it is enough to set only a few parameters, after which the mesh is automatically generated for a computational area that has the geometry of any complexity degree.

The calculation starts with some initial approximation set as the initial data. At each global iteration, the velocity and pressure fields, obtained after the time specified as the integration step, are calculated. The stationary solution of the problem is achieved after a sufficiently large number of global iterations corresponding to a long period of time.

When performing the calculation, the program automatically breaks the calculation area into separate subareas, subdomains, the size of which is proportional to the velocity of the appropriate processor. The calculation algorithm works in such a way that each processor calculates the flow only in its own subdomain, and after each global iteration, the resulting solution is “adhesion”. The parallelization algorithms used in modern software products let to achieve acceleration with an efficiency close to 100%. In other words, the calculation time is inversely proportional to the total computing power

As a rule, the adhesion condition is set as boundary conditions on all solid walls (the velocity is zero), the distribution of all velocity components in the inlet section, and the first derivatives (in the direction of flow) of the velocity components in the outlet section are equal to zero. In practice, if the flow rate at the inlet is approximately uniform in the intersection, the user sets only the average velocity (or flow rate) at the inlet, and sets nothing at the output - it is assumed that the user chose the output section far enough from the areas of intensive restructuring of the flow.

Examples of successful application of computer modeling for various processes of chemical technology are presented, in particular, in (Zhaofeng, 2006), (Liu et al., 2012), (Wang al., 2015), (Li and Li, 2016), Tahvildari et al., 2016).

MAIN FOCUS OF THE CHAPTER, RESULTS AND DISCUSSION

The qualitative picture of the typical regimes regarding the equipment operation is demonstrated below.

Vortex Granulator

Regime I. Gas flow filtering through the stationary layer of granules (fig. 1a). The layer is motionless, gas flow passes through pores in the layer.

Regime II. Beginning of the granules layer motion (fig. 1b). In this regime, the lower parts of the layer are partially weighted, while the upper parts of the layer are still motionless.

Regime III. Partially fluidized bed (fig. 1c). Granules in the lower parts of the layer are in the weighted state and move along the circular trajectory with low intensity, the granules in the upper parts of the layer are motionless. At a certain gas velocity, the whole layer goes into the weighted state.

Regime IV. Fluidized bed with partial swirling (fig. 1d). Granules in the weighted layer take the wave trajectory, which is characterized by the localized vortex motion, started in any random place of the layer. The vortex is spread in that part of the layer, which is above the gap of the gas-distributing device, and the rest of the area is in the suspended state. With an increase of the gas velocity, the growing number of granules is drawn into the vortex motion.

Regime V. Developed vortex fluidized bed (fig. 1e). With further increase of the gas velocity, the process of waves formation is weakened, the weighted layer has the ordered vortex structure.

Figure 1a. Typical regimes of the vortex granulator operation: gas flow filtering through stationary layer of granules

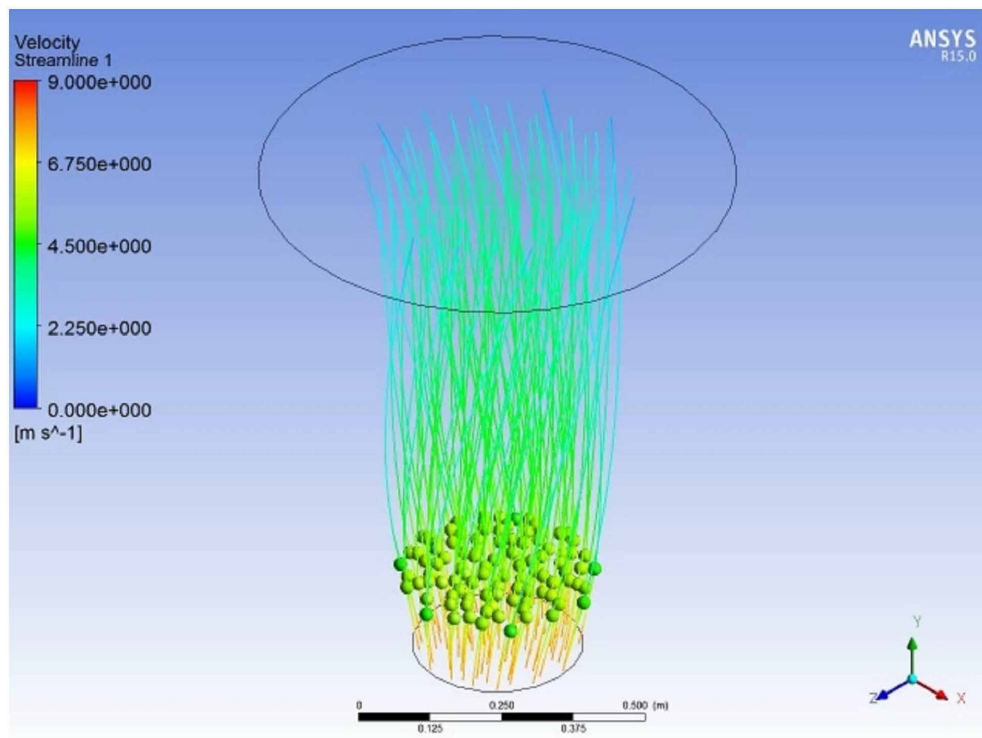


Figure 1b. Typical regimes of the vortex granulator operation: beginning of the granules layer motion

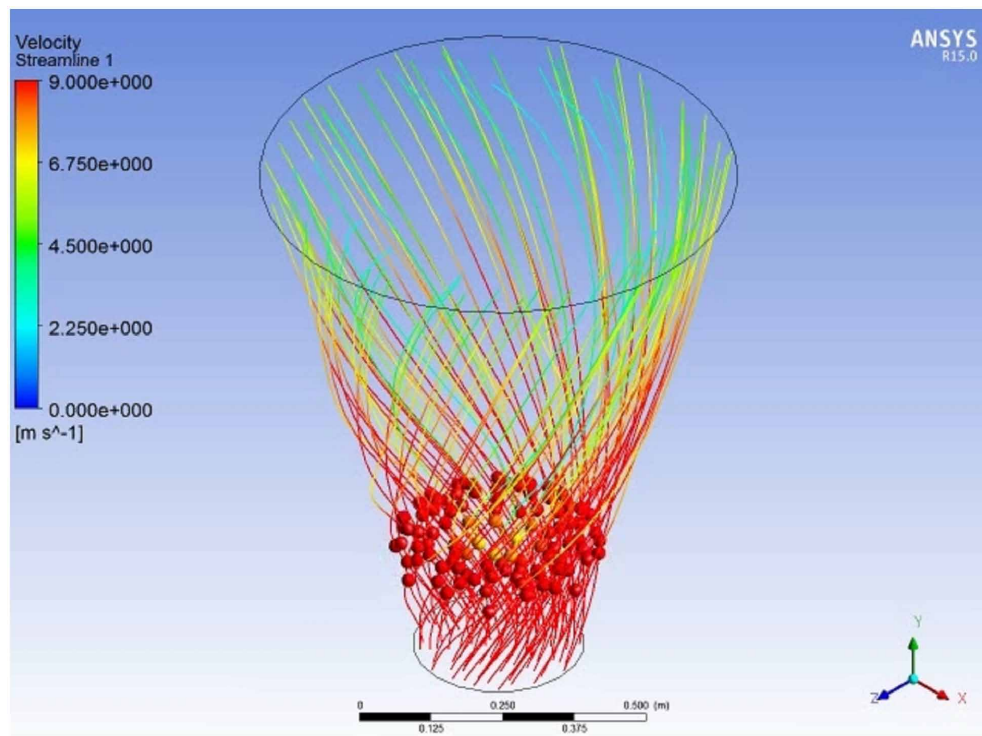


Figure 1c. Typical regimes of the vortex granulator operation: partially fluidized bed

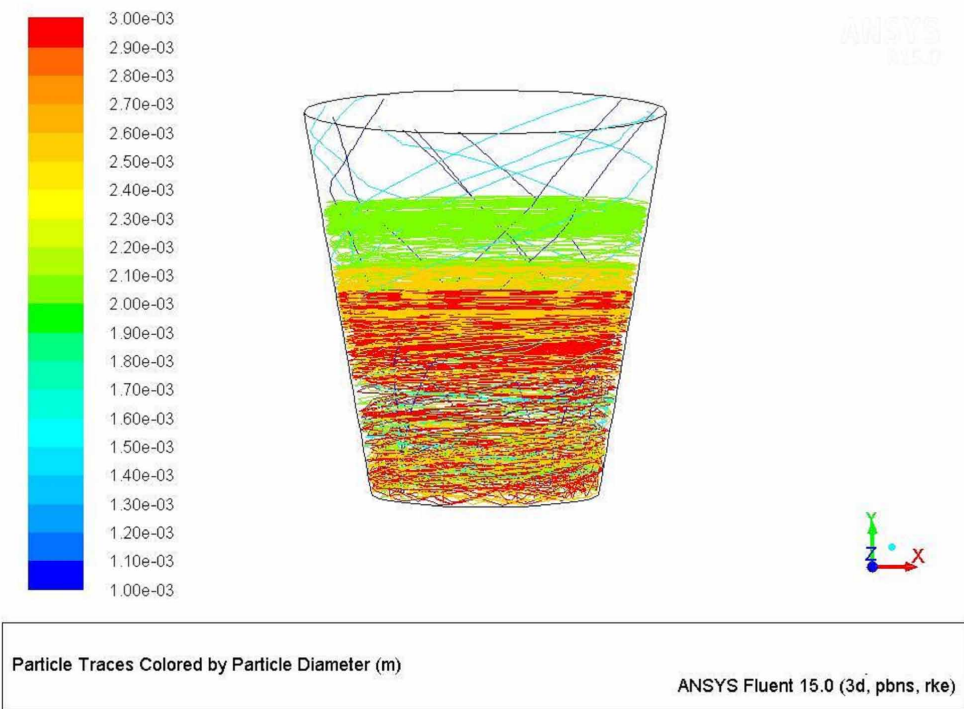


Figure 1d. Typical regimes of the vortex granulator operation: fluidized bed with partial swirling

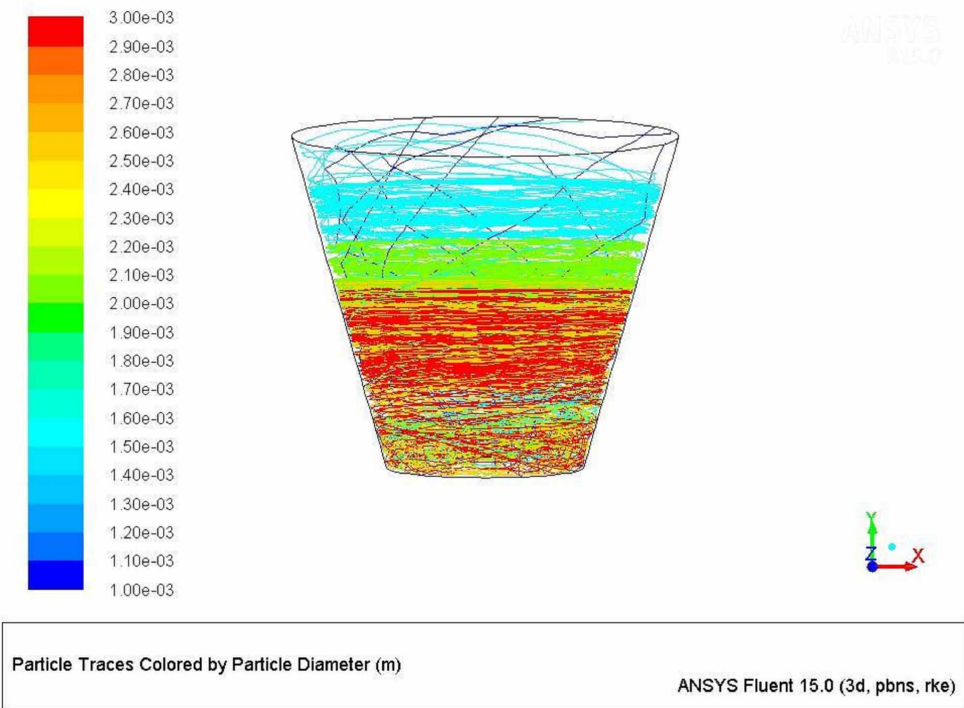


Figure 1e. Typical regimes of the vortex granulator operation: developed vortex fluidized bed

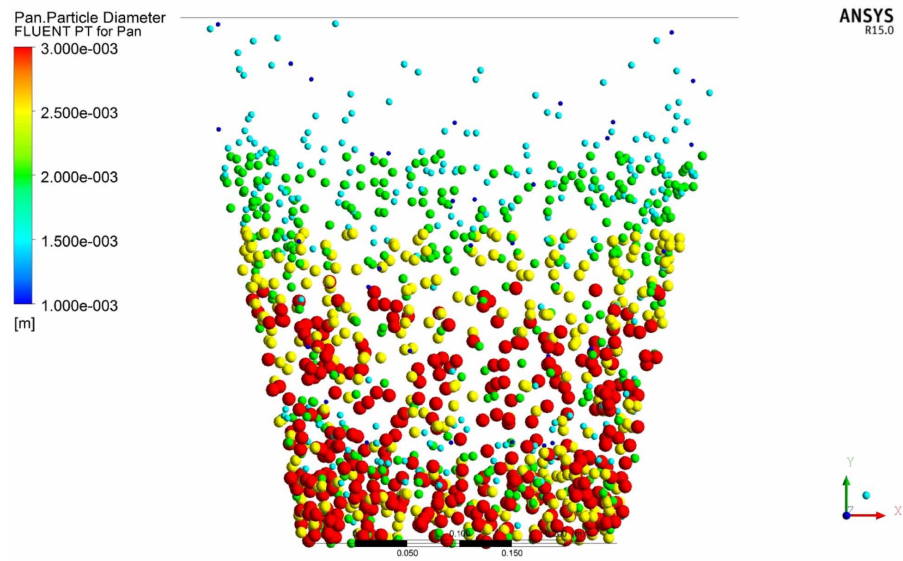
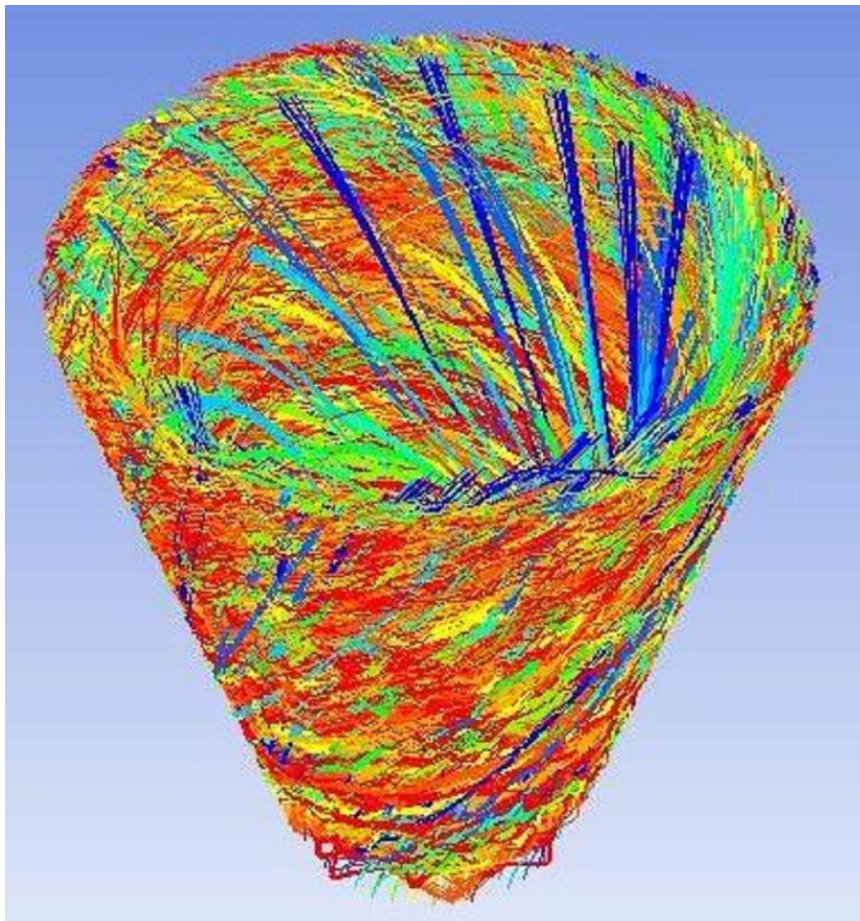


Figure 1f. Typical regimes of the vortex granulator operation: stable ablation regime of granules



Regime VI. Stable ablation regime of granules (fig. 1f). Space over the vortex weighted layer of the material is filled with the solid phase, and the difference between the vortex weighted layer and overlayer space consists only in the concentration of the solid material. There is a gradual reduction of the granules concentration in the overlayer space with an increase of the height. Granules leave the vortex layer; their ablation is caused by the domination of the upward flow gas force over the gravity force.

Shelf Multistage Device With a Weighted Layer

In the upper section of the device the inclined shelves (angle of lean 30°) with ratio l/B (where l is the distance from the end of the shelf to the wall of the device; B – width of the device) in the amount of $l/B=1$ (free channel) to $l/B=0.3$ were installed. The results of the previous investigations (Yukhymenko et al., 2016), show that the efficiency to extract the small fraction (with the size less than 1 mm) into the ablation has the maximum value with the ratio $l/B=0.5$.

Thus, at the first stage of theoretical studies, it was necessary to find the optimal value of the free area in the shelf for the ablation efficiency of the fraction less than 1mm.

The results of the computer modeling show that the efficiency to extract the small fraction into the ablation is achieved with the reduction of the perforation degree of the shelf. While installing the solid shelf in the separation zone of the device (fig. 7), the whole gas flow guaranteedly passes through the outloading space. It defines the extraction of the small fraction of the material precisely in this zone. 5% of perforation in the shelf causes the redistribution of the gas flow, directing its share through the hole in the shelf. Therefore, the gas flow intensively is interconnected with the material and not only in the outloading space zone but also on the shelf surface. It causes the increase of the fraction ablation efficiency, which is less than 1 mm.

The further increase of the free area of the inclined contact shelf is followed by the reduction of the small fraction extraction efficiency into the ablation. It is explained by the increased effect of the gas flow redistribution. The gas flow velocity in the outloading space decreases so much that the small fraction extraction process of the material in this zone is not very intensive. The total area of holes does not provide sufficient phase contact (Yukhymenko et al., 2016).

There are three perforated shelves in the drying (granulation) zone of the device. Moreover, while changing the distance between the end of the shelf and the wall of the device, it is possible to create different velocities of the gas flow

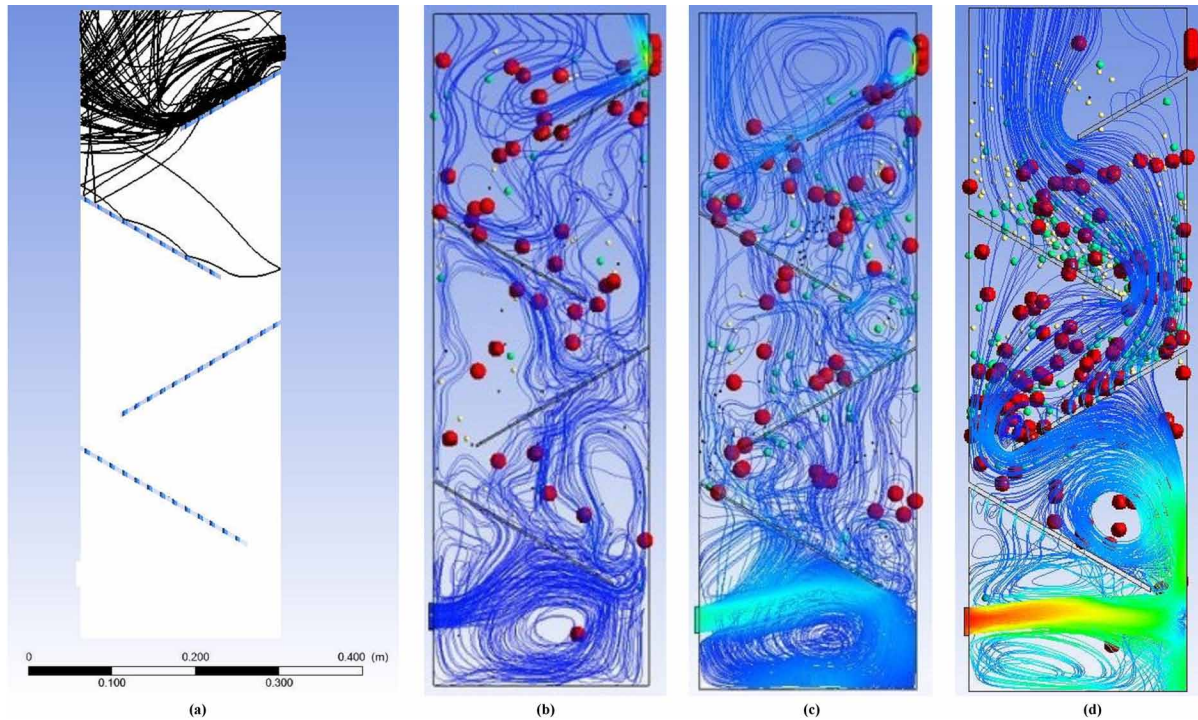
Regime I. Transition regime (figs 2a,b). The upward gas flow force in this regime causes the gradual change of its motion trajectory from the translational to the pulse-forward ones in direction of the outloading gap. Therefore the dispersed material starts to move into the weighted state, and the inertial force is compensated by the upward flow force of the drying agent.

Regime II. The weighted layer regime (fig. 2c). The effect, made by the upward flow of the heat transfer agent in such regime leads to the creation of the stable weighted layer thanks to the compensation of the inertial force and the rolling-down force on the sloping surface. The gas flow velocity reaches the first critical value, then it increases and is within the range of the working velocity.

Regime III. The regime of the developed weighted layer (fig. 2d). This regime is described by a dominant effect of the upward gas flow force on the dispersed material and by the increase of its displace-

ment vertical component. Therefore, at first, the share of the dispersed material is taken out from the surface of the shelf and then it moves to the outloading gap.

Figure 2. Typical operation regimes of the shelf multistage device (dryer or granulator): a – separation of the small fraction in the shelf dryer (granulator); b – transition operation regime of the dryer (granulator); c – weighted layer regime of the dispersed material motion; d – developed weighted layer regime of the dispersed material motion



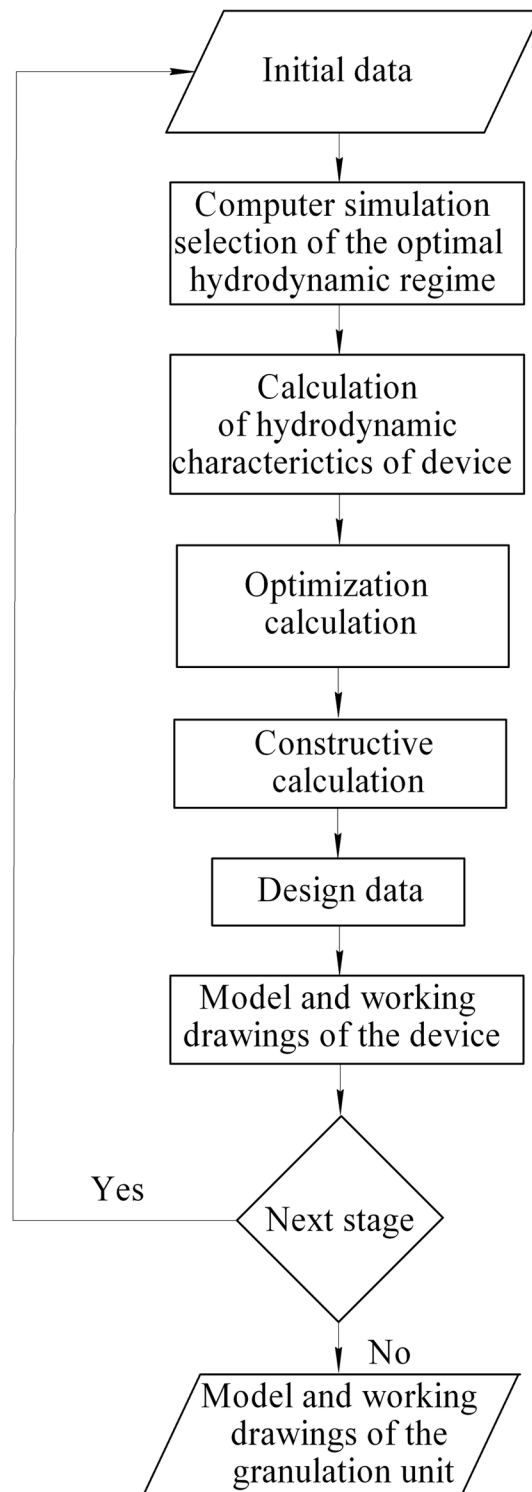
SOLUTIONS AND RECOMMENDATIONS

The above results for the modeling of the operating regimes in devices let to select the following optimal parameters to implement the target process:

- configuration of the workspace in the device and the way to create the directed motion of the dispersed phase flow;
- duration of the dispersed phase in the workspace of the device;
- technological features of the related processes (separation, classification, heat exchange)
- properties of the raw material and commodity product – temperature and humidity features, fractional composition, hardness index, special properties (porosity, selectivity etc.).
- energy indicators of the process implementation.

In general, the algorithm to calculate the device with the directed fluidized bed can be represented by means of blocks (fig 3), which are logically interconnected between each other. The peculiarity of the algorithm includes its universality and ability of “flexible” implementation of separate blocks.

Figure 3. Block-scheme of the granulation unit calculation



The modeling and calculation results with the use of authors' software, description of which is represented in the work (Artyukhov and Artyukhova, 2018), are exported to the cluster Granulation Unit[®] in the form of separate structural blocks (fig. 4).

Figure 4. Cluster Granulation Unit[®]

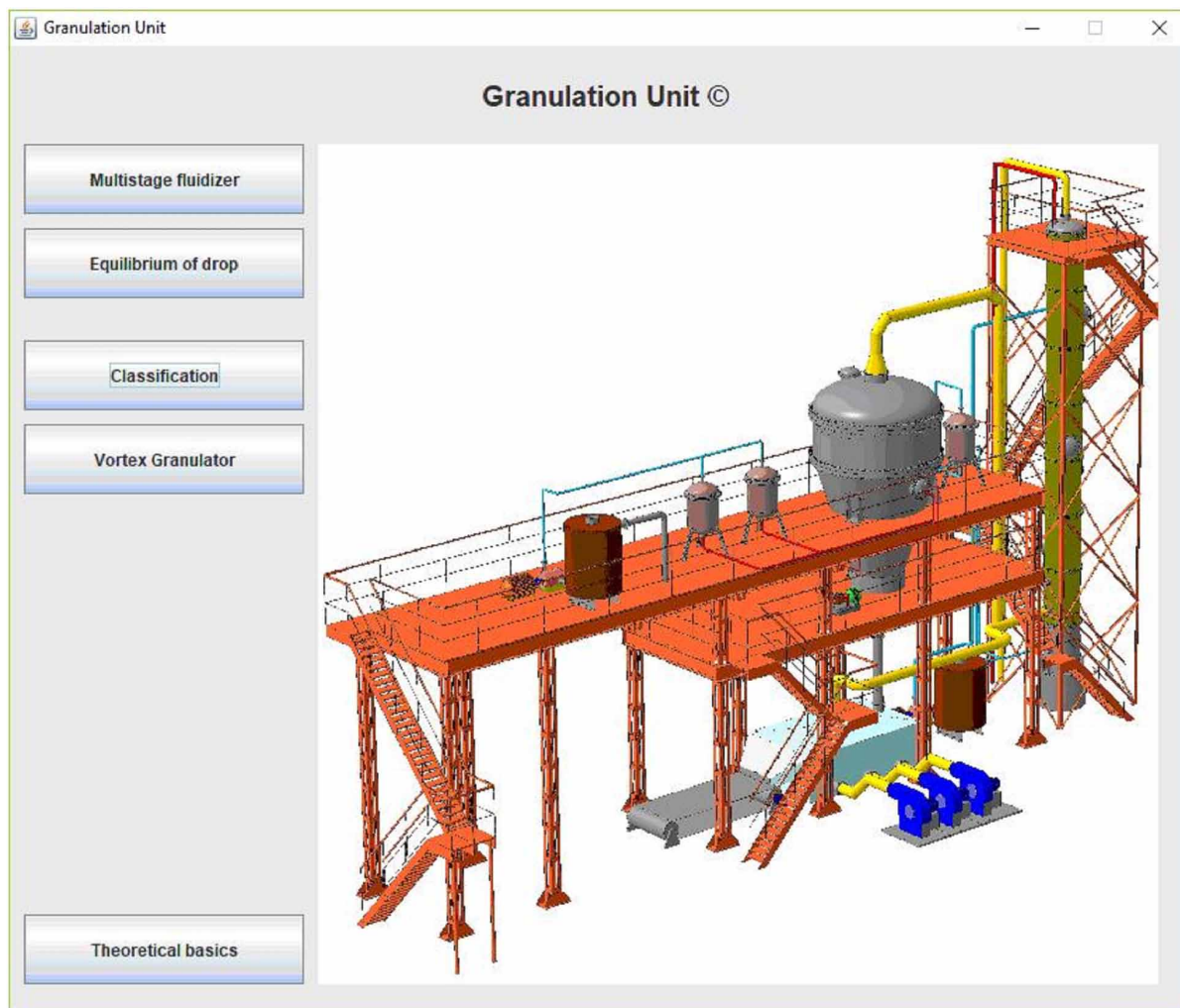
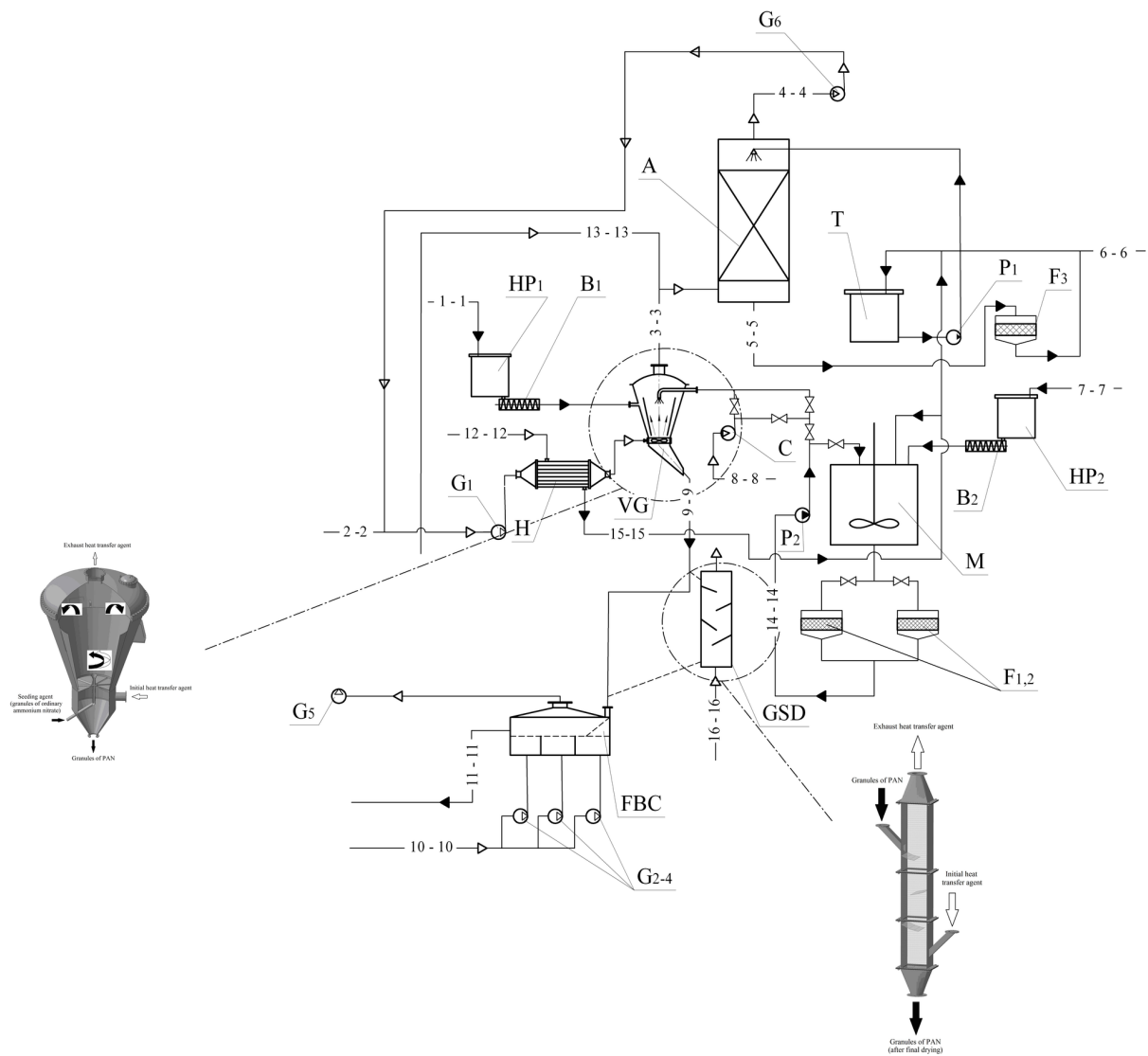


Fig. 5 demonstrates the technological scheme of the porous ammonium nitrate production. In this scheme the devices, operating regime modeling of which is shown above, are used as the main technological equipment.

Elements of the installation:

VG – a vortex granulator; H – a heater; GSD – a gravitational shelf dryer; FBC – a fluidized bed cooler; A – an absorber; F – a filter; M – a mixer; B – a batcher; HP – a hopper; G – a gas blower; P – a pump; T – a tank; C – a compressor.

Figure 5. A unit for the production of PAN granules by using a vortex granulator (multistage shelf granulator) and a gravitational shelf dryer (Artyukhov and Artyukhova, 2018)



The main flows:

1-1 – a seeding agent; 2-2 – manufacturing air; 3-3 – polluted air; 4-4 – purified air; 5-5 – polluted water; 6-6 – water; 7-7 – substandard granules; 8-8 – air for the spraying of the liquid materials (solution, melt); 9-9 – the product; 10-10 – air for cooling of granules; 11-11 – granules for packaging; 12-12 – steam; 13-13 – dusty gas; 14-14 – liquid materials (solution, melt); 15-15 – water condensate; 16-16 – a drying agent.

The operation principle of vortex granulator is the following.

A heat transfer agent is fed through the pipe into the device in the lower cylindrical section of the granulator. It flows to the vortex gas-distributing unit. The heat transfer agent is twisted around the vertical axis of the device and acquires a spiral motion. The vortex axisymmetrical flow of the heat transfer agent moves up the inner cone space towards the material. Simultaneously the fluid material (solution or fusion) is fed into the formed vortex flow of the heat transfer agent through the pipe in the upper part of the outing case into the spray unit. The fusion jet flows out of the spraying unit and is broken into the spheric granules. The created granules contact with axisymmetrical vortex flow of the heat transfer agent, get cool, are crystallized and come to the internal surface of the inner cone. Depending on the obtained size, granules are classified into the large and small fraction due to the whirl and axis velocity changes in the axisymmetrical vortex flow of the heat transfer agent heightwise of the inner cone. Small fraction granules are taken by the vortex flow of the heat transfer agent, move to the upper internal cone intersection and discharge from the active volume of the device through the annular space. In the annular space, the small fraction granules are moving down under the gravity force and get down to the lower intersection of the outing case. In the lower part of the outing case, these granules come to the depression zone. It is formed around the gas flow jet, which is fed through the pipe in the central part of the granulator (it is not shown on the scheme). Granules are sucked by this jet and are thrown to the central part of the inner cone workspace into the vortex weighted layer core. Fusion, that falls on the surface of the small granules, is crystallized. Therefore, the size of the granules increases. The large fraction does not leave the workspace of the device, and granules circulate in the inner cone and move down along its intersection, as far as they are growing and increasing. When the intended size is achieved, granules fall down on the surface of the inner cone, pass through the vortex gas-distributing unit and are output from the device. The spent heat transfer agent is driven from the inner cone through the proper pipe.

The operation principle of the gravitational shelf dryer is the following.

The drying agent, which at first is evenly distributed over the intersection of the body and then rises upwards, is brought to the lower section of the device. Therefore, it passes some (calculated, in this case – three) perforated inclined inserts (shelves), then it is output from the upper section of the device. Together with the motion of the drying agent, the dispersed material, which is necessary to dry, is fed to the upper shelf of the body. During the counteraction with the drying agent flow, the dispersed material is heated and the small fraction is removed with a purpose to balance the weighted layer porosity on the upper inclined contact shelf. The length of the inclined contact shelf provides the residence time of the dispersed material, which promotes the complete heating of the material, separation of the small dispersed particles and removal of the unbound humidity from the surface layer of the dispersed material during the constant drying velocity. The dispersed material is gradually moving along the upper shelf, is output through its free end and comes to the middle shelf, during drying by the drying agent flow. As the dispersed material moves through the middle shelf, unbound humidity is further removed from the surface layer of the dispersed material during the constant drying velocity as a result of the counteraction with the drying agent flow. The dispersed material is gradually output through the free end of the shelf and reaches the lower shelf. As the dispersed material moved through the lower shelf, the bound humidity is removed from the depth of the material during the drying velocity decrease. The dispersed material is output from the device after the drying process.

This device may be used to dry granular materials, for granulation, capsulation of the mineral fertilizers granules. In order to obtain the final product with a low level of the small dispersed (fine) particles, the process should be carried out with the initial separation, which is implemented on the upper shelf

of the device. There are hydrodynamic conditions under which the small fraction of particles (with the size less than 1 mm) is removed from the basic material.

FUTURE RESEARCH DIRECTIONS

The obtained results of the research have to be used in the modeling of the main technological equipment in the granulation unit under conditions of the granules variable mass and their compressed motion. For this purpose, it is necessary to implement the modeling of the granule heating kinetics, its dehydration and establishment of the gas flow motion intensity influence on the mass exchange processes.

The objective for further investigations is to create the algorithm to export the results of the constructive calculations of the main technological equipment in the granulation unit into the solid modeling programs.

After exportation one supposes to form the model of devices, design and technological documentation for their production.

After the implementation of the original idea to create compact and mobile granulation units through using the multiple-use technology of the drying agent's potential, a scientific method to calculate moving fluidized bed devices will be investigated by the possibility of utilizing harmful substances and by the heat and energy recuperation.

The obtained results of the project will let us define the complex effect made by the hydro- and thermodynamic conditions of the granulation process on the efficiency of moisture removal from dispersed particles with special properties (porous ammonium nitrate, encapsulated fertilizers with organic shell etc).

CONCLUSION

Based on the analysis of the ways to obtain the granulated product in units with implementation of the directed fluidized bed of the dispersed particles, it was defined that today there are no reliable mechanisms which control the residence time of the dispersed particle in the workspace of the device for such methods. It made impossible to obtain the granulated product with given qualitative indices with minimum energy costs for this process.

Cluster Granulation Unit® enables to carry out the optimization calculation of the main technological equipment in the granulation unit at the stage of its design without the expensive multifactorial experiment. The software peculiarities also let to carry out the equipment changeover if the raw material features and the productivity of the unit are changed.

The results of the optimization calculation with the software products and cluster implementation let to provide:

- a minimum “hydrodynamic” residence time of the particle in the device, which will not exceed “thermodynamic” time (minimum time of particles dehydration to the normative index is determined by the dehydration kinetics laws, thermodynamic indices of the dehydration process);
- implementation of different temperature and humidity potential of the heat transfer agent in some sections in the devices;
- variety of the device construction;
- possibility to use the recirculation of the heat transfer agent.

The original methods of the equipment engineering calculation with the directed fluidized bed as a part to obtain the granulated products are found.

The main recommendations to improve the technological equipment construction of the granulation unit have been implemented in the new methods of granulation, drying and classification in the directed fluidized bed and in devices for their fulfilment, protected by the patents.

ACKNOWLEDGMENT

This research work had been supported by the Cultural and Educational Grant Agency of the Slovak Republic (KEGA), project No. KEGA 002TnUAD-4/2019 and by the Ministry of Science and Education of Ukraine under the project “Technological bases of multistage convective drying in small-sized devices with utilization and heat recovery units”, project No. 0120U100476.

REFERENCES

- Agrawal, R., & Naveen, Y. (2011). Pharmaceutical Processing – A Review on Wet Granulation Technology. *International Journal of Pharmaceutical Frontier Research*, 1, 65–83.
- Artyukhov, A., & Artyukhova, N. (2018). Vortex and highly turbulent flows in the chemical engineering: a review of the specialized software products cluster to perform calculations for the granulation devices. In *Advances in Engineering Research* (Vol. 25, pp. 99–146). Nova Science Publishers.
- Artyukhov, A., Artyukhova, N., & Krmela, J. (2019). Computer simulation of the aerodisperse systems hydrodynamics in granulation and drying apparatus. In *Process Analysis, Design and Intensification in Microfluidics and Chemical Engineering* (pp. 277–321). IGI Global. doi:10.4018/978-1-5225-7138-4.ch010
- Artyukhov, A., & Sklabinskyi, V. (2015). Hydrodynamics of gas flow in small-sized vortex granulators in the production of nitrogen fertilizers. *Chemistry & Chemical Technology*, 9(3), 337–342. doi:10.23939/chcht09.03.337
- Artyukhov, A. E., Obodiak, V. K., Boiko, P. G., & Rossi, P. C. (2017). Computer modeling of hydrodynamic and heat-mass transfer processes in the vortex type granulation devices. *CEUR Workshop Proceedings*, 1844, 33–47.
- Artyukhov, A. E., & Sklabinskyi, V. I. (2013). Experimental and industrial implementation of porous ammonium nitrate producing process in vortex granulators. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, 6, 42–48.
- Caiyuan, Y., Tao, Q., & Xizhong, W. (2004). Heat and mass transfer in process of fluidized bed spray granulation. *Chinese Journal of Chemical Engineering*, 16(6), 836–839.
- Crowe, C. (2006). *Multiphase flow handbook*. Taylor & Francis Group.
- Gidaspow, D. (1994). *Multiphase flow and fluidization: continuum and kinetic theory descriptions with applications*. Academic Press.

Goldschmidt, M. J. V., Weijers, G. G. C., Boerefijn, R., & Kuipers, J. A. M. (2003). Discrete element modelling of fluidised bed spray granulation. *Powder Technology*, 138(1), 39–45. doi:10.1016/j.powtec.2003.08.045

Haron, N. S., Zakaria, J. H., & Mohideen Batcha, M. F. (2017). Recent advances in fluidized bed drying. *IOP Conference Series. Materials Science and Engineering*, 243, 012038. doi:10.1088/1757-899X/243/1/012038

Khanali, M., Rafiee, S., Jafari, A., & Banisharif, A. (2012). Study of Residence Time Distribution of Rough Rice in a Plug Flow Fluid Bed Dryer. *International Journal of Advanced Science and Technology*, 48, 103–114.

Li, X., & Li, G. (2016). Modeling and Simulation of Hydrodynamic Bubble-Liquid Turbulent Flows in Bubble-Column Reactors. *Chemical Engineering & Technology*, 39(12), 2380–2388. doi:10.1002/ceat.201600074

Liu, K., Gu, X. G., Ba, D. C., Li, P. Y., Du, G. Y., Yue, X. J., & Yang, N. (2012). Numerical Research on Flow Characteristics of Vortex Stage in Dry High Vacuum Pump. *Physics Procedia*, 32, 127–134. doi:10.1016/j.phpro.2012.03.529

Marchisio, D. L., & Fox, R. O. (2013). Computational Models for Polydisperse Particulate and Multi-phase Systems. Cambridge Series in Chemical Engineering. Cambridge University Press. doi:10.1017/CBO9781139016599

Ostroha, R., Yukhymenko, M., Litvinenko, A., Bocko, J., & Pavlenko, I. (2019). Granulation process of the organic suspension: Fluidized bed temperature influence on the kinetics of the granule formation. *Lecture Notes in Mechanical Engineering*, F2, 463–471. doi:10.1007/978-3-319-93587-4_48

Ostroha R., Yukhymenko M., Yakushko S. & Artyukhov A. (2017). Investigation of the kinetic laws affecting the organic suspension granulation in the fluidized bed. *Eastern-European Journal of Enterprise Technology*, 4/1(88), 4-10.

Palappan, K. G., & Sai, P. S. T. (2008). Studies on segregation of binary mixture of solids in a continuous fast fluidized bed: Effect of particle density. *Chemical Engineering Journal*, 138(1-3), 358–366. doi:10.1016/j.cej.2007.06.008

Patel, P., Telange, D., & Sharma, N. (2011). Comparison of Different Granulation Techniques for Lactose Monohydrate. *International Journal of Pharmaceutical Sciences and Drug Research*, 3, 222–225.

Philippsen, C. G., Vilela, A. C. F., & Zen, L. D. (2015). Fluidized bed modeling applied to the analysis of processes: Review and state of the art. *Journal of Materials Research and Technology*, 4(2), 208–216. doi:10.1016/j.jmrt.2014.10.018

Saikh, M. (2013). A technical note on granulation technology: A way to optimise granules. *International Journal of Pharmaceutical Sciences Review and Research*, 4, 55–67.

Sinaiski, E.G. (2010). *Hydromechanics: theory and fundamentals*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.

Tahvildari, K., Tavakoli, H., Mashayekhi, A., & Golmohammadzadeh, R. (2016). Modeling and simulation of membrane separation process using computational fluid dynamics. *Arabian Journal of Chemistry*, 9(1), 72–78. doi:10.1016/j.arabjc.2015.02.022

Trojan, M. (2015). Computer modeling of a convective steam superheater. *Archives of Thermodynamics*, 36(1), 125–137. doi:10.1515/aoter-2015-0009

Van Ommen, J. R., Valverde, J. M., & Pfeffer, R. (2012). Fluidization of nanopowders: A review. *Journal of Nanoparticle Research*, 14(3), 737. doi:10.1007/11051-012-0737-4 PMID:22593643

Wang, Q., Feng, Y., Lu, J., Yin, W., Yang, H., Witt, P. J., & Zhang, M. (2015). Numerical study of particle segregation in a coal beneficiation fluidized bed by a TFM–DEM hybrid model: Influence of coal particle size and density. *Chemical Engineering Journal*, 260, 240–257. doi:10.1016/j.cej.2014.08.052

Yukhymenko, M., Ostroha, R. & Artyukhov, A. (2016). Hydrodynamic and kinetic processes of the mineral fertilizer granules encapsulating in the multistage device with suspended layer. *Eastern-European Journal of Enterprise Technology*, 6(84), 22–28.

Zhang, H., Kong, W., Tan, T., Gilles, F., & Baeyens, J. (2017). Experiments support an improved model for particle transport in fluidized beds. *Scientific Reports*, 7(1), 10178. doi:10.1038/41598-017-10597-3 PMID:28860473

Zhaofeng, T. (2006). *Numerical modelling of turbulent gas-particle flow and its applications* (Ph.D. thesis). RMIT University.

Zimmermann, S., & Taghipour, F. (2005). CFD Modeling of the Hydrodynamics and Reaction Kinetics of FCC Fluidized-Bed Reactors. *Industrial & Engineering Chemistry Research*, 44(26), 9818–9827. doi:10.1021/ie050490+

ADDITIONAL READING

Ashcraft, R. W., Heynderickx, G. J., & Marin, G. B. (2012). Modeling fast biomass pyrolysis in a gas-solid vortex reactor. *Chemical Engineering Journal*, 207–208, 195–208. doi:10.1016/j.cej.2012.06.048

Bakshi, A., Altantzis, C., Bates, R. B., & Ghoniem, A. F. (2015). Eulerian–Eulerian simulation of dense solid–gas cylindrical fluidized beds: Impact of wall boundary condition and drag model on fluidization. *Powder Technology*, 277, 47–62. doi:10.1016/j.powtec.2015.02.056

Cano-Pleite, E., Hernandez-Jimenez, F., & Acosta-Iborra, A. (2015). Compressible-gas two-fluid modeling of isolated bubbles in a vertically vibrated fluidized bed and comparison with experiments. *Chemical Engineering Journal*, 271, 287–299. doi:10.1016/j.cej.2015.02.096

Chen, Q., Zhai, Z., & Wang, L. (2007). Computer modeling of multiscale fluid flow and heat and mass transfer in engineered spaces. *Chemical Engineering Science*, 62(13), 3580–3588. doi:10.1016/j.ces.2006.11.059

Hiltunen, K., Jäsberg, A., & Kallio, S. (2009). *Multiphase Flow Dynamics. Theory and Numerics*. Edita Prima Oy.

Kaewklum, R., Kuprianov, V. I., & Douglas, P. L. (2009). Hydrodynamics of air-sand flow in a conical swirling fluidized bed: A comparative study between tangential and axial air entries. *Energy Conversion and Management*, 50(12), 2999–3006. doi:10.1016/j.enconman.2009.07.019

Kudra, T., & Mujumdar, A. S. (2002). *Advanced Drying Technologies*. Marcel Dekker.

Lettieri, P., & Macrì, D. (2016). Effect of process conditions on fluidization. *Kona Powder and Particle Journal*, 33(0), 86–108. doi:10.14356/kona.2016017

Lipin, A. A., Nebukin, V. O., & Lipin, A. G. (2018). Simulation of heat and mass transfer processes under granules capsulation in fluidized bed. *Izvestiya Vysshikh Uchebnykh Zavedenii*, 61(4-5), 98–104. doi:10.6060/tcct.20186104-05.5624

Litster, J. & Ennis, B. (2004). *The science and engineering of granulation processes*. Springer-Science+Business Media.

Mazzei, L., Casillo, A., Lettieri, P., & Salatino, P. (2010). CFD simulations of segregating fluidized bidisperse mixtures of particles differing in size. *Chemical Engineering Journal*, 156(2), 432–445. doi:10.1016/j.cej.2009.11.003

Parise, M. R., Kurka, P. R. G., & Taranto, O. P. (2009). Identification of defluidization region in a gas-solid fluidized bed using a method based on pressure fluctuation measurements. *Brazilian Journal of Chemical Engineering*, 26(3), 537–543. doi:10.1590/S0104-66322009000300009

Salman, A. D., Hounslow, M. J., & Seville, J. P. K. (2006). *Granulation*. Elsevier Science Ltd.

Stahl, H. (2010). *Comparing Granulation Methods*. Hürth: GEA Pharma Systems.

Xianfeng, F., Zhufang, Y., & Parker, D. J. (2008). Prediction of bubble behaviour in fluidised beds based on solid motion and flow structure. *Chemical Engineering Journal*, 140(1-3), 358–369. doi:10.1016/j.cej.2007.10.013

Yang, W. C. (2003). *Handbook of fluidization and fluid-particle systems*. Marcel Dekker. doi:10.1201/9780203912744

KEY TERMS AND DEFINITIONS

Computational Model: A mathematical model in computational science that requires extensive computational resources to study the behaviour of a complex system by computer simulation.

Drying: A mass transfer process consisting of the removal of water or another solvent by evaporation from a solid.

Fluidised Bed: A physical phenomenon occurring when a quantity of a solid particulate substance (usually present in a holding vessel) is placed under appropriate conditions to cause a solid/fluid mixture to behave as a fluid.

Granulation: The act or process of forming into grains or granules.

Hydrodynamics: A branch of physics that deals with the motion of fluids (gases) and the forces acting on solid bodies immersed in fluids (gases) and in motion relative to them.

Navier-Stokes Equations: The equations that describe the motion of viscous fluid substance.

Reynolds-Averaged Navier-Stokes Equations: The time-averaged equations of motion for fluid flow.

Section 10

Environmental Science and Technologies

Renewable Resources and Value-Based Complex Forest Management

10

Yuri P. Pavlov <https://orcid.org/0000-0003-1927-5795>*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria*

INTRODUCTION

The human society is an inseparable part of nature and if we perceive it as complex and difficult to describe, this is a result of the capabilities of our consciousness. The needs of the society and the vital resources required for its existence are completely embedded in the aspects of the surrounding reality. The nowadays use of natural resources demands a more complex assessment approach that takes into account factors such as economic efficiency, social impact, environmental impact, etc. The manner in which we conceptualize and utilize these resources depends on the level of the collective and personal individual consciousness and its perfection. Different social groups perceive and use the natural resources with different aspect of intensity and different awareness. This determines the multiple aspects of the production processes in the process of objectives' multiattribute modeling and formation as well as in the control of these processes. Accounting for the social and group interests in the nature utilization processes have been analyzed by various contemporary authors (Farnsworth, 1983; Shukla & Dubey, 1997; Jungmeier, 2003).

In the literature are discussed different decisions making approaches, as the focus is on the different aspects of the group decision making as a response to the multiple aspects of the interests and needs of the various social strata representatives. In the considered approaches and methods it is sought to additionally account for the intensity of the social or group preferences and their combination in common relation. However, it is necessary to emphasize that measurement and comparison of preference intensities are mathematically permissible in the interval scale (Clark, 1990; Keeney & Raiffa, 1999; Jungmeier, 2003).

There are many investigations of the production processes and its modeling (Hotelling, 1931; Swallow, 1994, Clark, 1990). Widely known and discussed are the models for production management proposed and used by Canadian scientists Clark and Munro (Clark & Munro, 1975). Both types – the exhaustive and renewable production were considered. Clark's models aiming at optimal resource use have been applied to fishing industry in Canada and mining enterprise, but they can also be applied to other branches of the economy and human activity such as agriculture, mining, logging, ecology, etc (Clark, Clarke & Munro, 1979). As a leading scientist in the theory of non-smooth optimization, Clarke has made a brilliant analysis of the optimal control under such modeling and has determined a strategy for finding this control (Clarke, 1983). However, there remains the question of accounting for and including social, economic and other human factors as an extension to this modeling. In our view this can be achieved by using measurement theory and utility and multiattribute utility theory (Fishburn, 1970; Pfanzagl, 1971; Keeney & Raiffa, 1999).

In the present chapter we will consider value based control of logging process with jointly accounted for complex social, ecological and economic factors in the modeling and control. Cases of renewable natural resources are considered including economic efficiency and social effect for the population.

DOI: 10.4018/978-1-7998-3479-3.ch089

The new element is taking into account the social impact on the population, such as pay, ecological effects such as biodiversity and their combination with economic efficiency in logging. This is achieved by constructing a multifactor utility function which under the different control models is used as an objective function or as a part of the differential equation describing the production process. The utility function is constructed based on the preferences of biologist expert and reflects his attitudes with respect to the considered production – in this case logging (Lyubenova et al, 2015). Such utility function can be constructed for every stakeholder: government, business, ecologists, or all of them combined in a single multifactor utility function.

BACKGROUND

In ordinal or ranked scale it is possible to search for a common conforming relation only in the same scale. Under the same, in essence, ranked approach there appears the so called Arrow's paradox (Ekeland, 1983). For this class of problem and on that level of measurement the most appropriate mathematical solution is the Kemeny-Snell's median (Litvak, 1982). However, finding it requires "branch and bound" type algorithms and this makes these algorithms difficult to use in practice. If a higher scale for measurement of the preferences is taken, such as the interval, there still remains the question for the respective mathematical validity and availability of appropriate theory and implemented algorithms.

In our view, the most mathematically founded is the approach of utility theory and multiattribute utility theory (Keeney & Raiffa, 1999; Pavlov & Andreev, 2013). Their proper use permits the most complete inclusion of the mathematical and computational power in the process of multifactor modeling and the control of natural exhaustive or renewable resources. The measurement scale for the utility function is the interval scale which allows for exact assessment and comparisons of the expressed human preferences and their intensity with respect to the production process and different aspects of it (Fishburn, 1970; Pfanzagl, 1971). In fact, the utility function is an analytical representation of the preferences in the temperature or interval scale. From conceptual point of view, when assessing utility it is necessary to account for real processes such as inaccuracy in the expressed preferences, errors, information uncertainty, etc. All of that places higher requirements to the process of assessment and measurement of the preferences and their analytical representation as utility. Due to this reason it has been chosen to use the approach of measurement theory combined with the theory of stochastic programming for the assessment and practical implementation of the utility and preferences in practice (Pavlov, 2005; Pavlov & Andreev, 2013).

FOREST SYSTEM, PREFERENCES AND UTILITY EVALUATION

In preceding years, a multiattribute utility function has been built, taking into account three important factors pertaining to a forest system (Lyubenova et al, 2015). The main objective and structure of the study was formulated by a qualified bio ecologist expert. On the basis of his preferences, the corresponding single-attribute utility functions were evaluated, and using the structure of the problem and the type of independence by utility among the factors, sub-objectives a multiattribute utility was constructed. This multi-attribute utility function could be used in a value based production model as objective function (Keeney, 1988). It is a polynomial approximation of von Neumann's utility and mathematically this polynomial approximation provides the standard for the application of utility in optimal control problems.

The expert analysis carried out led to accepting the following sub-objectives and the appropriate criteria, which adequately describe the main objective and are real, physically measured quantities:

- X_1 -factor (material services as volume of timber per hectare- economic effect) [10-300 m³.ha⁻¹];
- X_2 - factor (supporting services - ecological effect) [1-200 number of species per hectare];
- X_3 - factor (percentage of employed locals in the forestry sector - social effect) [1- 30%].

Independence by utility was found between the following factors in the process of investigation:

- X_2 from X_1 ; X_2 from X_3 ;
- X_3 from X_1 ; X_3 from X_2 ;

Utility independence of X_2 from the other two factors means the following. The DM's preferences between lotteries for x_2 , $x_2 \in X_2$ at different values of x_1 , $x_1 \in X_1$ and x_3 , $x_3 \in X_3$ do not change, suggesting utility independence of x_2 from the changes of other two factors. Whatever the reserves of wood in the forest ecosystem are and the employment of the population in the forestry sector may be different, but in any case the preferences are directed to the presence of a large species richness of the forest, i.e. great variety of species of trees, grasses, moss, lichen etc. that form the ecosystem and ensure its greater stability.

The preferences for gambles of the decision maker (DM) for x_3 at different values of x_1 and x_2 do not change, suggesting utility independence of x_3 from the changes of other two factors (Farquhar, 1984; Keeney & Raiffa, 1999). This means that whatever the reserves of wood are and at different species richness of the forest ecosystem, in any case, the preferences are aimed at the increasing of number of workers in the forestry sector. Using the theory for decomposition of a multi-attribute utility to simpler functions the following multi-attribute utility structure is determined (Keeney & Raiffa, 1999):

$$U(x_1; x_2; x_3) = k_1 U(x_1; x_2^0; x_3^0) + f_2(x_1) \times [U(x_1^0; x_2; x_3^0)] + \\ + f_3(x_1) \times [U(x_1^0; x_2^0; x_3)] + f_{23}(x_1) \times [U(x_1^0; x_2; x_3^0)] \times [U(x_1^0; x_2^0; x_3)],$$

where $U(x_1^0; x_2^0; x_3^0) = 0$ and $U(x_1^*; x_2^*; x_3^*) = 1$.

In the formula above $X^0 = (x_1^0; x_2^0; x_3^0) = (10, 1, 1)$ and $X^* = (x_1^*; x_2^*; x_3^*) = (300, 200, 30)$. The functions f_2 , f_3 and f_{23} have the forms:

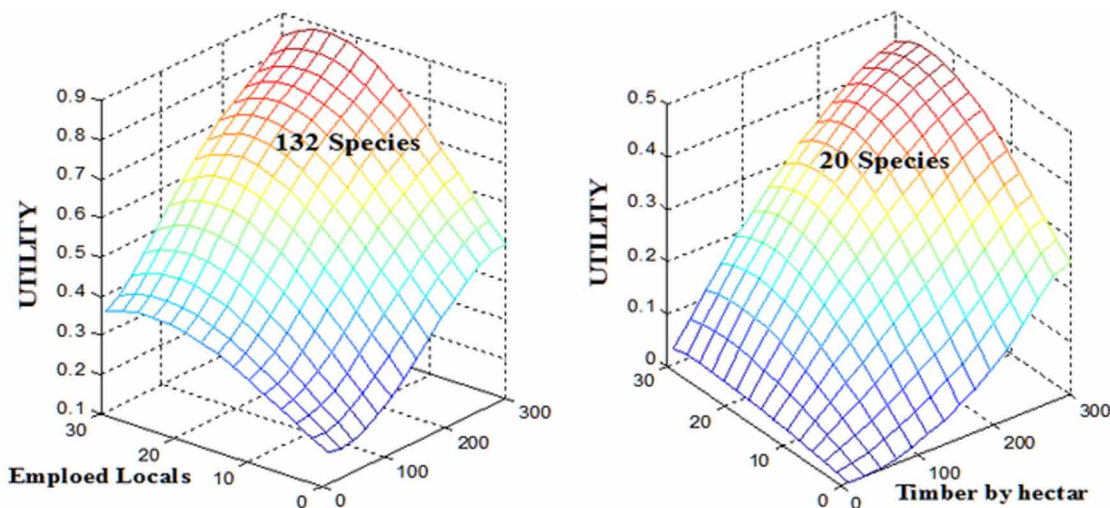
$$f_2(x_1) = U(x_1; x_2^*; x_3^0) - k_1 U(x_1; x_2^0; x_3^0), \\ f_3(x_1) = U(x_1; x_2^0; x_3^*) - k_1 U(x_1; x_2^0; x_3^0), \\ f_{23}(x_1) = U(x_1; x_2^*; x_3^*) - f_2(x_1) - U(x_1; x_2^0; x_3^*).$$

Each of these sixth functions was evaluated based on the DM's preferences. In the following figure (1) are shown utility functions $U(x_1; 132; x_3)$ and $U(x_1; 20; x_3)$.

The described value modeling is part of the Smart Forest Ecological Management System ForEco which is based on Open Data and FIWARE (Lyubenova et al, 2015).

For the sake of completeness, we will describe in short the algorithms for evaluating the utility functions u . Two sets are defined for stochastic machine learning approximation utility evaluation (Pavlov & Marinov, 2018).

Figure 1. Utility functions: A – 132 Species in the Forest; B – 20 Species in the Forest



$$A_u = \{(\alpha, x, y, z) / (\alpha u(x) + (1-\alpha)u(y)) > u(z)\},$$

$$B_u = \{(\alpha, x, y, z) / (\alpha u(x) + (1-\alpha)u(y)) < u(z)\}.$$

The approximation of the utility function is constructed by pattern recognition of the set A_u . The assessment process is a machine-learning approach based on the DM's preferences and the machine learning is a probabilistic pattern recognition procedure because $(A_u \cap B_u^c) \neq \emptyset$ and the utility evaluation is a stochastic approximation with noise (uncertainty) elimination (Pavlov & Andreev, 2013). The following presents the evaluation: DM compares the "lottery" $\langle x, y, \alpha \rangle$ with the simple alternative $z, z \in X$ ("better-", $f(x, y, z, \alpha) = 1$ ", "worse-", $f(x, y, z, \alpha) = -1$) or "can't answer or equivalent- $\sim$, $f(x, y, z, \alpha) = 0$ ". The discrete function $f(\cdot)$ denotes the qualitative DM's answer expressed as preference. This determine a learning, a training point $((x, y, z, \alpha), f(x, y, z, \alpha))$. The following recurrent stochastic algorithm constructs the utility polynomial approximation based on a finit set of learning points:

$$u(x) = \sum_i c_i \Phi_i(x),$$

$$c_i^{n+1} = c_i^n + \gamma_n \left[f(t^{n+1}) - \overline{(c^n, \Psi(t^{n+1}))} \right] \Psi_i(t^{n+1}),$$

$$\sum_n \gamma_n = +\infty, \sum_n \gamma_n^2 < +\infty, \forall n, \gamma_n > 0.$$

The notations are based on the structure of the set A_u : $t = (x, y, z, \alpha)$, $\psi_i(t) = \psi_i(x, y, z, \alpha) = \alpha \Phi_i(x) + (1-\alpha) \Phi_i(y) - \Phi_i(z)$, where $(\Phi_i(x))$ is a family of polynomials. The line above the scalar product $\bar{v} = (c^n, \Psi(t))$ means: $(\bar{v} = 1)$, if $(v > 1)$, $(\bar{v} = -1)$ if $(v < -1)$, and $\bar{v} = v$ if $(-1 < v < 1)$. The coefficients c_i^n take part in the polynomial representation:

$$g^n(x) = \sum_{i=1}^n c_i^n \Phi_i(x), (c^n, \Psi(t)) = \alpha g^n(x) + (1-\alpha)g^n(y) - g^n(z),$$

$(c^n, \Psi(t)) = G^n(x, y, z, \alpha)$, where $(c^n, \Psi(t))$ is a scalar product.

The mathematical procedure describes the following assessment process. The learning points could be set randomly or with a pseudo random sequence. The expert relates intuitively the “lottery” (x, y, z, α) to the set A_u with probability $D_1(x, y, z, \alpha)$ or to the set B_u with probability $D_2(x, y, z, \alpha)$.

$f(x, y, z, \alpha) = D_1(x, y, z, \alpha) + \zeta = 1$ if (x, y, z, α) is referred to A_u or

$f(x, y, z, \alpha) = D_2(x, y, z, \alpha) + \zeta = -1$ if (x, y, z, α) is referred to B_u .

The probabilities $D_1(x, y, z, \alpha)$ and $D_2(x, y, z, \alpha)$ are mathematical expectations of $f(\cdot)$ over A_u and B_u respectively:

$(D_1(x, y, z, \alpha) = M(f/x, y, z, \alpha))$ if $(M(f/x, y, z, \alpha) > 0)$,

$(D_2(x, y, z, \alpha) = (-)M(f/x, y, z, \alpha))$ if $(M(f/x, y, z, \alpha) < 0)$.

Let $D'(x, y, z, \alpha)$ is the random value:

$D'(x, y, z, \alpha) = D_1(x, y, z, \alpha)$ if $(M(f/x, y, z, \alpha) > 0)$;

$D'(x, y, z, \alpha) = (-D_2(x, y, z, \alpha))$ if $(M(f/x, y, z, \alpha) < 0)$;

$D'(x, y, z, \alpha) = 0$ if $(M(f/x, y, z, \alpha) = 0)$.

We approximate $D'(x, y, z, \alpha)$ by $G(x, y, z, \alpha) = (\alpha g(x) + (1-\alpha)g(y) - g(z))$. The coefficients c_i^n take part in the approximation of the function $G(x, y, z, \alpha)$ $G^n(x, y, z, \alpha) = (c^n, \Psi(t))$. The function $G^n(x, y, z, \alpha)$ is positive over A_u and negative over B_u depending on the degree of approximation of $D'(x, y, z, \alpha)$. The function $g_n(x)$ is the approximation of the utility function $u(\cdot)$ (Pavlov & Andreev, 2013). The utility $U(x, y, z)$ and the single attribute utility functions are normed between 0 and 1 and are measured in the interval scale. These facts permit determination of the coefficients involved in the multiattribute utility function (Keeney & Raiffa, 1999).

The coefficients in the utility formula are determined by comparisons of lotteries of the following type (Pavlov & Andreev, 2013):

$$\left\{ \begin{array}{c} U(x_1) \times \alpha \\ + \\ U(x_2) \times (1-\alpha) \end{array} \right\} \begin{array}{c} > \\ \approx \\ < \end{array} U(x_3).$$

With $U(\cdot)$ is noted the utility function. The values x_1 , x_2 and x_3 are fixed and is fulfilled the following $(x_1 \uparrow x_3 \uparrow x_2)$ [8, 14]. The questions to the DM are like lotteries in which we vary only the probability $\alpha \in [0, 1]$. The stochastic procedure is of the type of Robins-Monro but the convergence is much quicker.

The procedure is the following. Let's α is a uniformly distributed random value in $[0, 1]$. We define the following random vector $\chi = (\eta_1, \eta_2, \eta_3)$, where:

- a) If $\{\}$ $\Rightarrow \chi = (\eta_1 = 1, \eta_2 = 0, \eta_3 = 0)$;
- b) If $\{\}$ $\Rightarrow \chi = (\eta_1 = 0, \eta_2 = 0, \eta_3 = 1)$;
- c) If indiscernibility $(\approx) \Rightarrow \chi = (\eta_1 = 0, \eta_2 = 1, \eta_3 = 0)$.

Let χ^n is a learning sequence of independent random values with equal to χ distribution. The stochastic recurrent procedure is the following:

$$(\lambda_1^{n+1}, \lambda_2^{n+1}, \lambda_3^{n+1}) = \Pr_P [(\lambda_1^n, \lambda_2^n, \lambda_3^n) + \gamma_n ((\eta_1^n, \eta_2^n, \eta_3^n) - (\lambda_1^n, \lambda_2^n, \lambda_3^n))],$$

$$\sum_1^\infty \gamma_n = \infty, \quad \sum_1^\infty \gamma_n^2 < \infty, \quad \gamma_n \geq 0, \forall n \in N, \text{ for example } \gamma_n = \frac{1}{(n+1)} \text{ with initial point is } 0.$$

The notation $\Pr_P$ has the meaning of projection over the set:

$$P = \{(\lambda_1, \lambda_2, \lambda_3) / \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0, \lambda_1 + \lambda_2 + \lambda_3 = 1\}.$$

The searched value is determined in the end as the following:

$$U(x_3) = U(x_2) + (\lambda_1 + \lambda_2/2)(U(x_1) - U(x_2)).$$

The completed stochastically utility evaluation theory is exposed in the monographs (Pavlov & Andreev, 2013).

RENOUVABLE PRODUCTION MODELING AND EFFECTIVE FOREST MANAGEMENT

As noted in the introduction, the Clark and Munro models will be analyzed (Clark & Munro, 1975; Clark, Clarke & Munro, 1979). These models are used in the Canadian fishing industry and have been discussed multiple times by other scientists. The model in an abridged version has the form:

$$\begin{aligned} \dot{y} &= F(y) - qyK, \\ F(y) &= r(y(\theta - y)). \end{aligned}$$

In the above formula q , r , and θ are constants ($q=0.7$, $r=1.1$, $\theta=1$). With (y) is denoted the available renewable natural resource (timber per hectare- X_1), and with K we mention the funds invested in the logging (employed locals in the forestry sector- X_3). Such models are used and could be used in other areas of activity where the resources are renewable (Jungmeier, 2003). It is obvious that equilibrium in the forest system occurs when the derivative of the quantity of wood (the derivative of y) is zero. In the

Clark's model the optimization criterion is the maximum profit. This does not, however, take into account the social impact of production on the population as well as the ecological conditions at the moment. These factors could be reported by including in an integral (objective function) a multi-attribute utility function and directed it at the maximum. The goal is to maximize the integral:

$$J = \int_0^T U(y, K)(py - c)K dt .$$

In this formula $U(y, k)$ is the utility $U(x_1; 132; x_3)$, (p) and (c) are constants ($p=1.3$, $c=0.31$), and T is the end of the production period. The constant (p) reflects the market price of the wood and the constant (c) reflects its procurement cost. The problem thus defined is an optimal control problem. The control input are the money- the fund K invested in the wood industry.

In the Clark-Munro's model in place of utility is introduced an exponential function, which varies with time. This reflects the devaluation of capital in the time. A constant could be set for shorter periods of investigation. The inclusion of the utility function as a quantity under the integral allows the social and environmental factors to be accounted for by it. When choosing an optimal solution about the production, these factors will influence the solution depending on the type of the utility function. A solution based on the principle of maximum of Pontryagin may be sought (Pontryagin et al, 1964; Alekseev et al, 1979; Clarke, 1983; Gabasov & Kirilova 1981). The necessary conditions for optimality following the Pontryagin's maximum principle are the arguments that maximize the Hamiltonian:

$$H(T, \psi, y_{opt}, K_{opt}) = \max_k [\psi(F(y) - qyK) + U(y, K)(py - c)K], \quad 0 \leq K \leq K_{max} .$$

With $\Psi(\cdot)$ is noted an auxiliary covector-function $\psi(t)$, subject to the adjoint (linear) differential equation:

$$\dot{\psi} = - \frac{\partial H(T, \psi, y_{opt}, K_{opt})}{\partial y} .$$

For autonomous systems, as in this paper, is performed

$$\max_K H(T, \psi, y, K) = h, \quad h \text{ is a constant.}$$

The type of optimal control is determined by the following manifold:

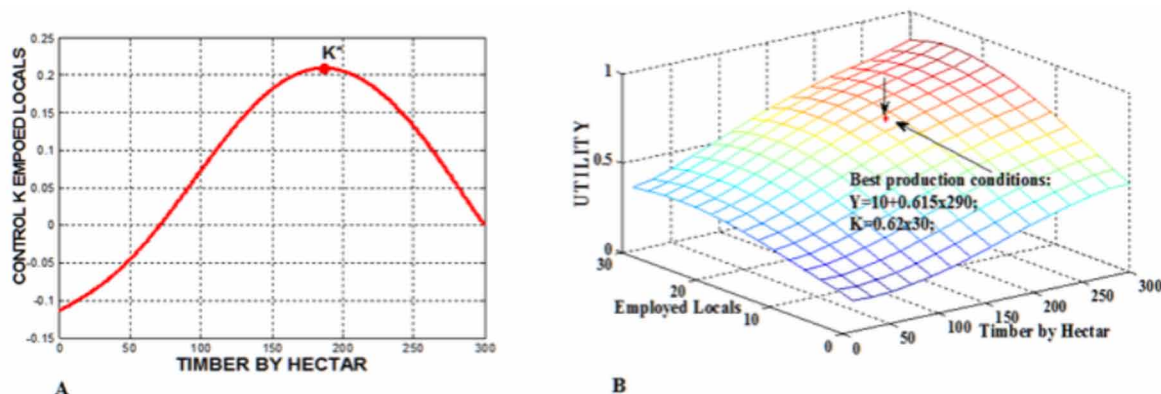
$$M(y, K, \psi) = (U(y, K)(py - c) - \psi qy)K = 0.$$

The optimal control K is be determined by the manifold as follows:

$$M(y, K, \psi) > 0 \Rightarrow K = K_{max} ,$$

$$M(y, K, \psi) < 0 \Rightarrow K = 0.$$

Figure 2. Optimal production conditions: A – Equilibrium point; B – Objective Utility function



Thus defined the control problem defines “chattering control” or “bang-bang control”. Obviously in the logging for a period of several years this is unacceptable. It is impossible for instance to have six months of logging and subsequently six months of total forest recovery. Such management could be acceptable in fisheries or in agriculture for some renewable resources. A possibility is to explore the manifold $M(y, K, \psi) = 0$ and determine the renewable natural resource y as function of the input K and hence the derivative of y as function of K , derivative of K and derivative of y . Conversely, this will allow determination of the control K as a function of y and the derivative of y . Of course, if a solution is found. This control guarantees a sliding mode on the manifold. Such a sliding on the manifold will move the system to an optimal equilibrium point or an optimal periodic mode, depending on the type of the used utility function. The fed-back control could permit the autonomous system to remain on the manifold. The beginning point will determine the moment, crosses the manifold. After this moment the system will stay on the manifold. All of the discussed above is possible only for autonomous systems and without taking into account the effect of capital devaluation.

However, the availability of a multi-attribute utility function allows another possibility, another “good” solution with simpler calculations. In the literary sources are discussed steady states that are optimal for economic reasons (Clark, Clarke & Munro, 1979). We can approach the problem in the same way and look for a state of equilibrium that determines the maximum of the integral criterion J . The maximum of the function under the integral determines an equilibrium point as steady state possibility:

$$\begin{aligned}\Phi(y) &= U(y, K)(py - c)K \\ &= U(y, \frac{F(y)}{qy})(py - c)\frac{F(y)}{qy}.\end{aligned}$$

In the expression above the control K is replaced by a formula which is determined by the differential equation which described the renewable natural resource y assuming the derivative zero. The function $\Phi(y)$ and the optimal point (y^*, K^*) are shown in Figure 2a. In figure 2b is shown the utility function $U(x_1; 132; x_2)$ and the equilibrium point (y^*, K^*) . The location of this optimal point is marked with stars.

The differential equation that describe the dynamic of y determines the minimal time to attain the equilibrium, starting from any initial position (y_n, K_n) . If y_n is less than y^* ($y_n < y^*$), it is obviously necessary to let nature return to state y^* . This determines $K=0$. Conversely, if y_n is greater than y^* ($y_n >$

y^*), then the control input is maintained maximum ($K = K_{\max}$) until the equilibrium point (y^*, K^*) is reached. In this manner, good production parameters can be identified which take into account the complex social and environmental factors in logging. We need to clarify that this is not the mathematically accurate optimal control. The optimal solution generally requires a mathematical study in details of the manifold $M(y, K, \psi) = 0$ described above.

We can approach this problem in a different way in order to avoid the chattering control or the exploration of the sliding manifold. The optimization problem could be described as a Meyer's problem (Clarke, 1983; Alekseev et al, 1979; Gabasov & Kirilova 1981). For this purpose, we determine the control K by the differential equation which described the renewable natural resource y and its derivative:

$$K = \Phi(y, \dot{y}) = \frac{F(y) - \dot{y}}{qy},$$

where $F(y) = ry(\theta - y)$.

In the optimal control problem the initial conditions are known and are as follows: $y: [0, T] \rightarrow \mathbf{R}$, with initial conditions $y(0)$ and $\dot{y}(0)$. These two initial conditions determine exactly the control K . It is need one additional variable s , $s(0) = 0$ and it is additionally set a new restriction:

$$\phi(y, s, \dot{y}, \dot{s}) = -\dot{s} - U(y, \Phi(y, \dot{y}))(py - c)\Phi(y, \dot{y}) \leq 0.$$

The above formula includes the utility function and we underline once again that the utility function in this form takes into account the economic and social effect of the state of the forest system in a fixed biodiversity.

A new optimization problem will be defined. The new objective functional is $f(s, y) = s$. The optimization problem is reformulated as minimization of $f(s(T), y(T)) = s(T)$, subject to the above restriction:

$$\begin{aligned} \min s(T), \\ \dot{y} &= F(y) - qyK, \\ F(y) &= r(y(\theta - y)), \\ \phi(y, s, \dot{y}, \dot{s}) &\leq 0. \end{aligned}$$

Careful analysis shows that the optimal problem so defined is equivalent to the initial optimization problem. The equivalent mathematical formulations are as follows:

$\max J,$

$$\text{where } J = \int_0^T U(y, \frac{F(y) - \dot{y}}{qy})(py - c)\Phi(y, \dot{y})dt,$$

$$\dot{y} = F(y) - qyK,$$

$$F(y) = r(y(\theta - y)),$$

$$\phi(y, s, \dot{y}, \dot{s}) \leq 0.$$

The following mathematical result is used for the solution of this optimal problem. Let the continuous by parts curve $y(\cdot)$ is a weak local minimum in Meyer's problem and satisfy the regularity conditions described in the Clarke's monograph (Clarke, 1983). Then there is a curve $p(t)$, a measurable function $\lambda: [0, T] \rightarrow \mathbb{R}$ and a constant λ_0 equal to 0 or 1 for which the following is satisfied:

$$\begin{bmatrix} \dot{p}(t) \\ p(t) \end{bmatrix} \in \lambda(t) \partial \phi(y, s, \dot{y}, \dot{s}) \text{ almost everywhere (a.e.);}$$

$$\lambda(t) \leq 0 \text{ a.e., } \lambda(t) = 0 \text{ when } \phi(y, s, \dot{y}, \dot{s}) < 0;$$

$$-p(T) \in \lambda_0 \partial f(s(T), y(T));$$

$$\|p(\cdot)\| + \lambda_0 \neq 0.$$

The mathematical expressions for $p(t)$ and its derivative $\dot{p}(t)$ are determined with the use of the above mathematical result. The time derivative of $p(t)$ is compared with the already determined mathematical expression for $\dot{p}(t)$. This comparison permits the determination of the second derivative $\ddot{y}$ as a function of y and $\dot{y}$. In this way, the derivate $\dot{K}$ of the control input K is determined at any moment. Then the optimal control is could be obtained as a solution of the following differential equation:

$$\dot{y} = r(y(\theta - y)) - qyK,$$

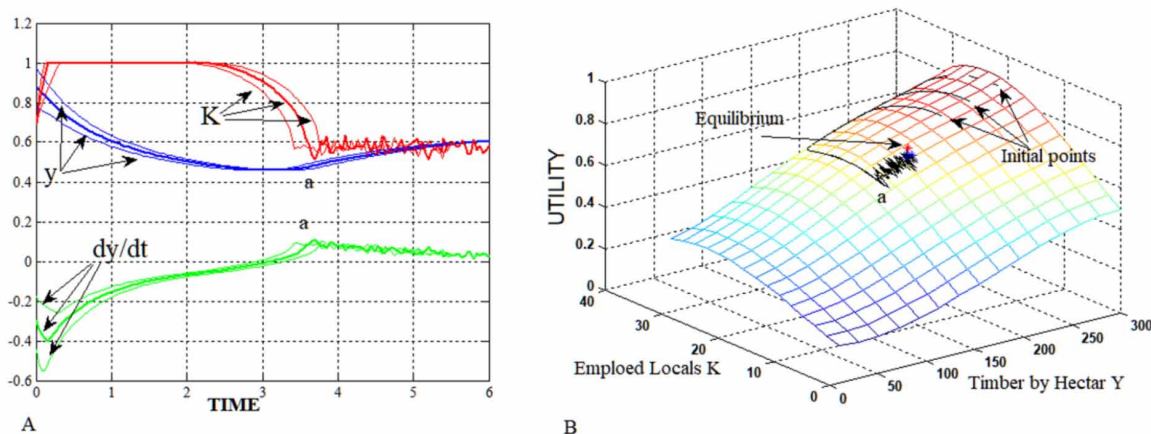
$$\dot{K} = -\frac{\ddot{y}}{qy} - \frac{r\dot{y}}{y} + \frac{1}{q} \left(\frac{\dot{y}}{y} \right)^2,$$

$$0 \leq K \leq K_{\max}.$$

For completeness of the exposition we will give the formula for the second derivative $\ddot{y}$. In the formulas below, we mark with U the utility function:

$$U(y, K) = U(y, \frac{F(y) - \dot{y}}{qy}).$$

Figure 3. Optimal production control: A – Optimal control; B – Optimal trajectories in connection with Utility function



The second derivative is described by the following formula:

$$B(y, \dot{y}) = \left[\frac{\partial U}{\partial y} \frac{F(y)}{qy} + \frac{\partial U}{\partial K} \left(\frac{-r}{q} \right) \frac{F(y)}{qy} - U \frac{p}{(c - py)} \frac{F(y)}{qy} + \frac{\partial U}{\partial K} \frac{F(y)}{q^2 y^3} \dot{y} + U \left(\frac{-r}{q} \right) \right],$$

$$\Gamma(y, \dot{y}) = \left[\frac{\partial^2 U}{\partial K \partial y} \cdot \frac{(-1)}{qy} + \frac{\partial^2 U}{\partial K^2} \frac{r}{q^2 y} - \frac{\partial U}{\partial K} \frac{p}{(c - py)} \frac{(-1)}{qy} + \frac{\partial U}{\partial K} \frac{1}{qy^2} \right] \dot{y} + \frac{\partial^2 U}{\partial K^2} \left(\frac{-1}{q^2 y^3} \right) (\dot{y})^2,$$

$$\Theta(y) = \left[\frac{\partial^2 U}{\partial K^2} + \frac{\partial U}{\partial K} \right] \frac{1}{(qy)^2},$$

$$\ddot{y} = \frac{B(y, \dot{y}) - \Gamma(y, \dot{y})}{\Theta(y)}.$$

It is necessary to be careful with the choice of the differential solver due to the peculiarities of Meyer's problem. At low values of the derivatives, the calculations are slow. The second derivative formulation of $\ddot{y}$ is quite large and heavy. But the discussed decision procedure allows anyone interested to find them. Only standard derivative calculations and enthusiasm are required. It is emphasized that the utility function $U(y; 132; K)$, has a polynomial mathematical description and it participates in these mathematical formulations, including with its own partial derivatives on the corresponding variables. In figure 3a are shown three optimal control solutions starting from three different initial points. In the figure 3b are shown the optimal trajectories of these processes in connection with the utility function.

It is obvious that the time minimization control is close to the optimal control for some initial conditions. In this manner, good production parameters can be identified which take into account the complex social and environmental factors in logging.

This solution focuses attention on complex models and mathematically well-founded methods where the decision maker is represented through his preference as a value or utility function, being part of the

mathematical modeling process. In fact, it is a value-based design, an engineering strategy grounded on system analyses enabling multidisciplinary design optimization (Keeney, 1988; Collopy & Hollings, 2009).

DISCUSSIONS

We underline that in the place of the utility function is putted a constant in Clark and Munro's models. Under these conditions, the optimal problem cannot be defined as Meyer's optimal problem, due to some particularities of the renewable production model. For this case is made an excellent analyze by Clarke (Clark, Clarke & Munro, 1979; Clarke, 1983). The demonstrated solution is possible because of the inclusion of the analytic utility function and the possibilities to calculate the partial derivatives.

It is seen that the DM's preferences representation as polynomial utility function allows being used the full power of optimal control theory in its modern form. Utilization of multiattribute utility functions in complex problems, taking into account complex social, economic, environmental, ecological and other factors permits a new value based modeling, value driven control and management synchronous with DM's preferences (Keeney, 1988; Ekeland, 1979). For some complex systems and problems, this is the only way for a good mathematical description.

These solutions allow good planning of logging processes when selecting the relevant forest parameters and wood prices. Whatever the semantic interpretation of the discussed models, the determination of optimal control follows one and the same mathematical procedures. That is to say, the model and utility representation can be set different semantics in different production conditions, but the mathematical procedure and solution remains the same. We repeat, this is partly due to the polynomial utility approximation.

CONCLUSION

The rational actions of the DM in complex forest processes are of great importance for good production results. The qualitative nature of man's evaluation and analysis is a limitation to this process. This imposes the need of developing specific approaches and methods for the analysis and evaluation of qualitative, conceptual information. Human values (utility) are the internal motivation for determining the goal and find out a good solution. One possible approach to the problem is the multiattribute utility approach. In that aspect, the program of actions of the DM towards the problem is a hierarchical decomposition of the considered goal. Preferences are expressed concerning the chosen goal and sub-objectives and value and utility functions are made that are later used in the standard mathematical processing.

The constructed multiattribute utility function reflects the DM preferences and attitudes toward the three forest factors, ecology, economic effect and social effect of the logging. In the renewable production model and exhaustible production model the utility function could be included in two different ways. The utility function is a part of the objective function in the renewable model, while in the exhaustible production model the utility function could be part of a differential equation. Such a value based modeling permits determination of optimal control and value driven management strategy in line with the DM's preferences regarding the multifactor problem.

In conclusion, we would like to point out that the model is actually a complex mathematical model of the kind "human-dynamic production", taking into account the complex social and economic characteristics of the timber production process.

REFERENCES

- Alekseev, V., Tihomirov, V., & Fomin, S. (1979). *Optimal Control*. Nauka.
- Clark, C. W. (1990). *Mathematical Bioeconomics: The Optimal Management of Renewable Resource* (2nd ed.). John Wiley & Sons.
- Clark, C. W., Clarke, F. H., & Munro, G. (1979). The Optimal Management of Renewable Resource Stocks: Problems of Irreversible Investment. *Econometrica*, 47(1), 25–47. doi:10.2307/1912344
- Clarke, F. (1983). *Optimization and non-smooth analysis*. New York: Wiley.
- Clark, C. W., and G. Munro. (1975). “The Economics of Fishing and Modern Capital Theory: A Simplified Approach. *Journal of Environmental Economics and Management*, 2(2), 92–106. doi:10.1016/0095-0696(75)90002-9
- Collopy, P., & Hollingsworth, P. (2009). Value-driven design. In *Host publication of Ninth AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*. Hilton Head, SC: American Institute of Aeronautics and Astronautics. 10.2514/6.2009-7099
- Ekeland, I. (1979). *Elements d'économie mathématique*. Mir.
- Farnsworth, M. C. (1983). The social impact of forest development in Northland. *New Zealand Journal of Forestry*, 28(2), 247–254.
- Farquhar, P. (1984). Utility Assessment Methods. *Management Science*, 30(11), 1283–1300. doi:10.1287/mnsc.30.11.1283
- Fishburn, P. (1970). *Utility theory for decision-making*. Wiley. doi:10.21236/AD0708563
- Gabasov, R., & Kirilova, F. M. (1981). *Methods of Optimization*. Nauka.
- Hotelling, H. (1931). The economics of exhaustible resources. *Journal of Political Economy*, 39(2), 137–175. doi:10.1086/254195
- Jungmeier, G. (2003). System Analysis of Forestry, Forest products and Recovered Wood. *International Conference: Efficient Use of Biomass for Greenhouse Gas Mitigation*.
- Keeney, R., & Raiffa, H. (1999). *Decision with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.
- Keeney, R. L. (1988). Value-driven Expert Systems for Decision Support. *Decision Support Systems*, 4(4), 405–412. doi:10.1016/0167-9236(88)90003-6
- Litvak, B. (1982). *Expert information, analyses and methods*. Nauka. (in Russian)
- Lyubenova, M., Pavlov, Y., Chikalanov, A., Spassov, K., Novakova, G., & Nikolov, R. (2015). Smart forest ecological management system. *Journal of Balkan Ecology*, 18(4), 363–375.
- Pavlov, Y. (2005). Subjective preferences, values and decisions: Stochastic Approximation Approach. *Proceedings of Bulgarian Academy of Sciences*, 58(4), 367–372.
- Pavlov, Y., & Andreev, R. (2013). *Decision control, management, and support in adaptive and complex systems: Quantitative models*. IGI Global. doi:10.4018/978-1-4666-2967-7

Pavlov, Y., & Marinov, E. (2018). *Value based modeling and control: preferences, utility, stochastic control*. Lambert Academic Publishing.

Pfanzagl, J. (1971). *Theory of Measurement*. Physical-Verlag. doi:10.1007/978-3-662-41488-0

Pontryagin, L., Boltyanskii, V., Gamkrelidze, R., & Mishchenko, E. (1964). The mathematical theory of optimal processes (D. E. Brown, Trans.). The Macmillan Co.

Shukla, J. B., & Dubey, B. (1997). Modelling the depletion and conservation of forestry resources: Effects of population and pollution. *Journal of Mathematical Biology*, 36(1), 71–94. doi:10.1007/002850050091

Swallow, K. S. (1994). Foundation Renewable and Nonrenewable Resource Theory Applied to Coastal Agriculture, Forest, Wetland, and Fishery Linkages. *Marine Resource Economics*, 9(4), 291–310. doi:10.1086/mre.9.4.42629089

KEY TERMS AND DEFINITIONS

Complexity: A condition of a system or situation integrated with some degree of order but with too many elements and relationships to be understood in a simple analytic or logical way. In the extreme, the complex system or situation is with multiple and diverse connections with dynamic and interdependent relationships, events, and processes.

Dynamic Model: Represents the behavior of an object over time. Dynamical systems theory is an area of mathematics used to describe the behavior of the complex dynamical systems, usually by employing differential equations or difference equations. When differential equations are employed, the theory is called continuous dynamical systems.

Model-Driven Decision-Making and Control: An emphasized access to and manipulation of a statistical, financial, optimization, or simulation model. It uses data and parameters provided by users to assist decision process in analyzing a situation.

Optimal Control: Is the process of determining *control* and state trajectories for a dynamic system over a period of time to minimize a performance index.

Utility Independence: An attribute A1 is utility-independent of attribute A2, if conditional preferences on lotteries on A1, given at a fixed value of A2, do not depend on that fixed point. The utility independence is not symmetrical.

Utility Theory: A normative approach to the issue of how people should rationally choose in conditions of uncertainty.

Value-Driven Design: A system engineering strategy, which enables multidisciplinary design optimization. It creates an environment for optimization by providing designers with a value function as objective or as part of the mathematical model.

Agriculture 4.0 and Bioeconomy: Strategies of the European Union and Germany to Promote the Agricultural Sector – Opportunities and Strains of Digitization and the Use of Bio-Based Innovations

10

Immo H. Wernicke

German Government (retired), Germany

INTRODUCTION

Agriculture 4.0 and Bioeconomy are new strategies of the European Union and the German Government for promoting Research and Development and the implementation of digital technologies and bio-based innovations in the agricultural sector. The programs upgrade the traditional Common Agricultural Policy and enlarge the legal and institutional framework of the *Acquis Communautaire*. Millions of agricultural holdings in Europe are encouraged to use biomass, renewable energies, digital innovations, and high tech machinery provided under the label of Industrie 4.0. The discussed strategies aim to achieve sustainable growth and competitiveness in agriculture, to secure self-sufficiency in food and energy production in Europe and to cover the Sustainable Development Goals of the United Nations. Both concepts provide opportunities to overcome the increasing demand for food by an exploding population. To be discussed are risks due to high investment costs, cyber crime, and poor public funds for improving the fragmented digital and energetic infrastructure. The improving of the official database is a precondition to get more information that is accurate on the impact of Agriculture 4.0 and Bioeconomy.

BACKGROUND

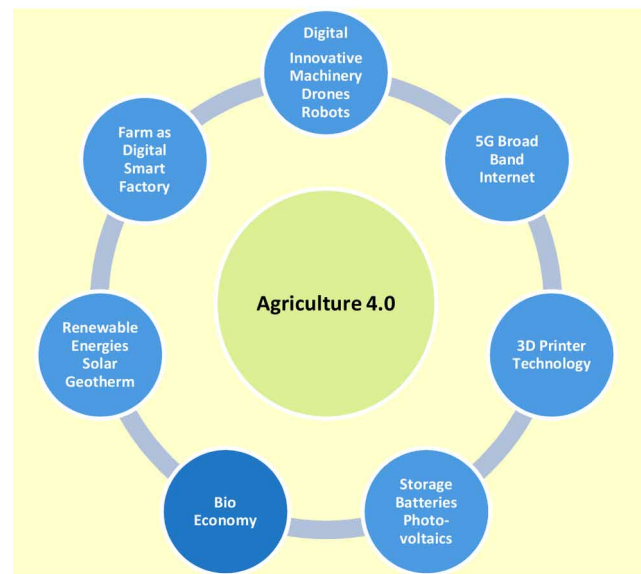
The term **Agriculture 4.0** is used as an official label of the 2020 strategic program of the European Commission (EC) and the German Government for promoting the agricultural sector, rural areas and the nutrition industries to achieve sustainable growth (EU Commission, 2018 a).

The term **Bioeconomy** is defined by the Directorate-General for Research and Innovation of the EC as a new economic system based on the sustainable use of renewable biological resources such as crops, forests, fish, animals and micro-organisms to produce food, materials and energy (European Commission, 2017).

Agriculture 4.0 is an innovative program for the benefit of the agricultural sector that refers to the Industrie 4.0 strategy of the EC and the German Government to foster the digitization of the Manufacturing Industries (Price Waterhouse, 2015). It goes back to the idea of the advisor to German Chancellor Angela Merkel Jeremy Rifkin (2015), who recommends a Digital Revolution in Europe also in the agricultural sector (Rossner, E. 2015). Digital innovations from the Information and Communications Technologies (ICT) sector including Start-ups and high tech machinery from the Manufacturing Industries are to be implemented in the complete supply chain from animal production and the use agricultural land to food

DOI: 10.4018/978-1-7998-3479-3.ch090

Figure 1. Basic Technologies of the Agriculture 4.0 concept



Source: Overview from different reports on Agriculture 4.0 by author

processing and distribution and finally to the selling to the consumers on cellular phone demand. **Figure 1** presents the main technologies of Agriculture 4.0.

The European Commission's Science and Knowledge Service (2016) being founded to evaluate the innovation program of the EC approves "New technologies and their adoption by EU farmers are key drivers in maintaining European agriculture competitive in a global world." Agriculture 4.0 is also connected to the Bioeconomy program that was launched by the EC in 2012 (EU Commission, Directorate-General for Research and Innovation, 2017). The EC Program Biotechnology and Life Sciences precedes the Bioeconomy concept as it was set up in 1982. The EC evaluates its former Biotechnology Strategy as "an important step towards a competitive and sustainable Knowledge Based Bio-Economy (KBBE)." The new strategy includes the traditional concept of Biotechnology, such as genes technology and all biotechnological applications across sectors in primary production, in industry and in the health sector described in detail by Patermann, C. and Aguilar, A. (2018).

As new components of the complex European and German institutional and legal system the Bioeconomy- and Agriculture 4.0-strategies have led to a multifold set of parliamentary decisions, regulations, funding and procurement programs, internet platforms, reviews, reports, staff working papers, working groups, manuals and other publications. The article focuses on the presentation and discussion of a few selected activities.

Generally said, these new strategies in agricultural politics supplement the already existing gigantic legal and institutional framework in Germany and in the EU laid down in the *Acquis Communautaire* that consists of about 100.000 pages covering the full scope of European legal issues. The new agricultural strategies of the EU and Germany are going well beyond the traditional Common Agricultural Policy (CAP), which is already providing about 50 Billion € as subsidies to the farmers in Europe every year and for decades (EU Commission, 2014 b).

Millions of agricultural holdings and family farms are being encouraged by the new concepts to make use of technical innovations, Bioeconomy, and renewable energies to achieve growth and competitive-

ness in agricultural production. The strategies are also addressed to academics, technical universities, and research institutes for receiving official funds for Research and Development (R&D) activities on these issues.

The concepts and their implied digital technologies have spread worldwide and influenced international agricultural politics, the media, and academic research. They became basic components of international fairs and put as topics of the agendas of international conferences, including the G 20 and the G7 meetings. Both concepts are complying with the recommendations of the FAO (2015), the Food and Agricultural Organization of the United Nations, and with the Sustainable Development Goals of the United Nations (2015). They seem to become a promising innovative agricultural policy also in developing countries for achieving worldwide sustainable growth in agriculture, forestry, in the nutrition industries and in rural areas also to tackle climate change and to overcome the increasing demand for food due to the increasing of world population. The FAO (2015, 2017) predicts about 9 billion people will need food and water in 2030.

OPPORTUNITIES AND STRAINS OF AGRICULTURE 4.0 AND BIOECONOMY

Opportunities for the Academia and Research Institutes

The new strategies provide many opportunities for the academia and research institutes. The EU Commission and the German Ministry of Education and Research (2011, 2014) are funding scientific institutes to implement the transformation from an oil-based to a bio-based industry. They also foster academic research on bio-based innovations for achieving sustainable economic growth in agriculture.

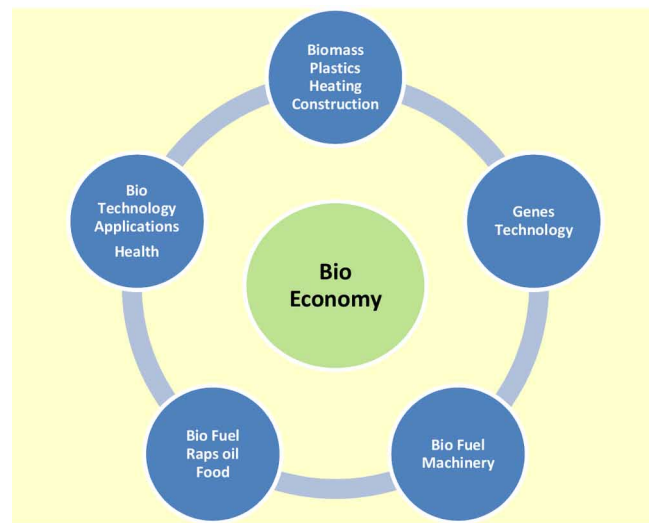
One of the new research projects is labeled as BioKompass. The Fraunhofer Institute Germany (2018) is charged to implement the project and to explore the transformation of the economy into a Bioeconomy. The German Research Center Juelich (2018) plays a key role, too, by using new techniques such as terrestrial systems for research on crop and plant production, and the Universities of Hohenheim and Vechta are specialized on master theses on these issues.

International High-level Summits on Agriculture and Bioeconomy in Berlin

The traditional international fair on agriculture in Berlin focuses many of its activities on both strategies. The Agricultural Ministry and the German Farmers Association (2015) have established a yearly international high-level conference on agricultural politics and put Bioeconomy as a topic on the agenda of the International Agriculture Ministers' Summit in Berlin in 2015. The German Minister of Food and Agriculture Schmidt (2015) initiated an international resolution, he accentuated "Bioeconomy will represent an opportunity for everyone and agricultural policy must go hand in hand with coherent economic, trade and energy policies - this applies to both national and global policies." Agricultural Ministers of 70 States and delegates of the FAO, the World Bank, and the EU-Commission have passed a resolution on Bioeconomy later put on the agenda of the G 20 and the G7 conferences.

The German Government has founded the German Bioeconomy Council (2018) that plays a key role in organizing the Global Bioeconomy Summits in Berlin: "Bioeconomy promotes distinct features of interest to several SDG (Sustainable Development Goals of the United Nations), including knowledge-based growth and jobs, the renewability of resources, resilience of ecosystems...". This international forum is a platform for information exchange between international experts and for discussing new policy strategies and initiatives from a global perspective. **Figure 2** presents the main technologies of Bioeconomy.

Figure 2. The new Bio economy Strategy and its Technologies



Source: Overview from different reports on Agriculture 4.0 by author

European Innovation Partnership for Agricultural Productivity and Sustainability

On the European level, the EU Commission launched the European Innovation Partnership for Agricultural productivity and Sustainability (EIP-AGRI) in 2012, preceding the two new strategies. It aims to foster a competitive and sustainable agriculture and forestry sector that “achieves more from less.” The EIP-AGRI contributes to ensuring a steady supply of food, feed, and biomaterials and a sustainable management of the essential natural resources on which farming and forestry depends. It brings together innovative farmers, holders, advisors, academics, researchers, and business managers and helps to build bridges between research and practice.

The European Technology Platforms (ETP) on Bioeconomy and Biomass

The EU Commission (2018 b) has also set up nine European Technology Platforms (ETPs) to foster Bioeconomy and biomass. The result of the collaboration between experts from the academia, industry, civil society organizations, farmers and forest owners have been published in the White Paper on the European Bioeconomy 2030. The White Paper of the EU Commission (2018 c) covers a large scope of issues like biomass, bio energy, and renewable energies like solar and wind energy. A basic technology outlined in the White Paper refers to the sustainable production and conversion of biomass into a variety of agricultural, industrial, and energetic products: food, health, fiber etc.

Renewable biomass encompasses any biological material (agriculture, forestry, and animal-based including fish) are to be used as a product in itself or as raw material. The White Paper also provides policy recommendations and covers a wide range of agricultural sectors – arable and livestock farming, forestry, food, aquaculture, the chemical industry, materials manufacturing and energy – to create a smart, sustainable, and inclusive European Bioeconomy.

The experts of the FAO, the OECD (2015) and the EU Commission (2018 d) have identified biomass as a prior growth factor in agriculture, as it is usable for the production of food, bio-plastics, transport fuels, electricity, heat, chemicals, and medicines.

The International Fair Agritechnica in Hanover Promotes Agriculture 4.0

The Agritechnica (2017), the international fair for agricultural machinery and industrial innovations in Hanover in Germany, functions as a key promoter of Agriculture 4.0 and Bioeconomy. The world's leading trade fair on farm machinery has become an international platform for the presentation of new technological developments in agriculture and for promoting negotiations between industrial exhibitors and managers of agricultural holdings and farmers. The German Agricultural Association (2015, 2017, and 2018) is one of the organizers of the fair that is presenting a wide range of innovations and technologies from the agricultural engineering industry and the Industrie 4.0 program, such as mobile machinery, hydraulic systems, engines, axles and drive technology, cabs, electronic systems, and ICT software. It offers a broad technical program including special shows, discussion forums inviting exhibitors, engineers, ICT experts, holding managers and farmers to discuss specific topics for future plant production.

The fair exhibits technological products, experience, and expertise on smart technologies to support farmers in practicing smart farming for a “green future.” Smart technologies are serving to reduce investment costs, allow better organizing specific jobs, reducing energy consumption, serving to protect soil, optimize nutrient and water use, and assisting to identify and treat pests and weeds. (Agritechnica, 2017)

Technologies of Agriculture 4.0 provided by Industrie 4.0

The EU Commission's Science and Knowledge Service (2016) accentuates that “New technologies and their adoption by EU farmers are key drivers in maintaining European agriculture competitive in a global world.” The European strategic concepts are backed up by large governmental funds and the public budget. The means focus support on the micro level to farms and farmers, cooperatives and S+M companies. (EU Commission, 2016)

The ongoing technical progress in the industrial world supplies high-tech agricultural machinery and equipment, such as drones and unmanned helicopters and seems to approve the vision of Rifkin, J. (2015) of a future digital world based on renewable energy. The nutrition industries and the logistics sector will also benefit from the new Cyber Physical System, which “stands for a demand driven decentralized and intelligent production process.” The digitizing of the transport and storing system in the complete supply chain might overcome the crucial challenge of managing waste of agricultural products and food. (McClelland, J. 2015)

Robotics, Drones and Unmanned Helicopters, Radio Controlled Agricultural Machinery

Robotics and Robots belong to the basic technologies of the German Industrie 4.0 – strategy presented at the International Fair on Industries Hanover (2015, 2018). Robots can take over many activities traditionally done by the farmers or their workforce. Digitization also allows a large scale seeding. German farmers interviewed by the Boston Consulting Group (2015) perform harvesting of fragmented farmland of small holders by setting up “virtually consolidated” plots of small farmers across field boundaries.

Drones and unmanned helicopters, first developed for military use, are useful innovative technical instruments for the agricultural sector. Being equipped with digital cameras and sensor techniques drones and unmanned helicopters enable the monitoring of the surface of arable land and the geological conditions of soil. The use of drones and unmanned helicopters allows a continuous monitoring of fruit growing and crops plantations, too. As part of digitized early warning systems, drones and unmanned helicopters can reveal threats on crops and can provide more accurate and reliable estimates of future yield of harvests. (Van Vark, C. 2016)

These digital technologies enable farmers for a more efficient and precise use of seeds, fertilizer and pesticides for crop plantation. **Precision farming** has become a prior topic of international discussion and negotiations. In January 2019 at the International Agricultural Ministers Conference in Berlin the German Agricultural Minister, Klöckner, J. (2019) has promoted the digitization of the agricultural sector internationally: “The decision to establish an international Digital Council is a breakthrough!” **Figure 3** presents digital equipment for precision farming.

Figure 3. Digital equipment for precision farming at International Green Week 2019



Source: Photo taken at International Green Week Berlin Germany 2019

Radio controlled agricultural high tech machinery such as harvesters run by bio fuel or solar energy make seeding, plowing and harvesting more efficient. Livestock farmers and holdings can profit from digitally controlled automatic feeding and milking machines for cattle. (Bailey P. 2013)

Smart Technologies for use by Smaller Holdings in Agriculture

Smart technologies for the crop production process enable to better organize labor input and energy consumption, they save costs, help protect the soil, optimize nutrient and water use, or help to identify and treat pests and weeds more efficiently. The fair Agritechnica (2017) presents also different smart technologies at each stage of the crop-production process to improve working conditions of farmers. Tillage technologies are combining tillage with fertilizer application (side dressing), strip till will combine deep tillage with no-till, and better straw management is enabled after harvest. Seeding technologies are combining seeding and fertilizing, and site-specific seed densities. Fertilizer technologies improve fertilizer planning, soil testing, and enable sensor-based fertilizing, technologies on plant protection like sensors, drones and satellites enable improved forecasts, pest detection and the reduction of spray rates and a single-plant treatment. Harvest technologies including sensors, drones, and satellites can reduce harvest losses, detect optimal harvest times, improve logistics, and better residue management. Software and data management systems provide field catalogues combined with forecast models or decision support models. (AGCO, 2019)

Farms and Agricultural Holdings as Smart Factories

Digital innovations in container systems developed by Industrie 4.0 provide farmers new opportunities to set up “mobile intelligent factories” for the processing of food. Both the agricultural sector and the nutrition industries might profit from the new container system for more mobility in farming and food processing. Additive Manufacturing allows an ICT supported small-scale production by the use of the 3D-Printer-Technology. All pulverized solid materials, metals, and plastics are producible by biomass, and they can be processed by the 3D-Printer-Technology. Farmers will be enabled to produce their technical equipment and to spare parts of their agricultural machinery in containers at their premises. Internet access provided by broadband techniques or mobile satellite systems will improve the farmer’s management activities for marketing and sales promotion (Agritechnica, 2017).

The digital revolution enables farmers to become shareholders of companies of the nutrition industries or of transport and logistics companies. The performance of the nutrition industries can benefit significantly through the digitization of food processing, too. The establishment of sharing models and the building up of cooperatives make these innovative techniques also affordable to small holders.

Renewable Energies Provide Opportunities for Farmers and Agricultural Holdings

The German Ministry for Economic Affairs and Energy (2015) and the European Union have passed their new Strategy 2020 on renewable energies. The strategy promotes the transformation from nuclear and fossil energies to a mix of the full scope of renewable energies, such as wind energy, hydropower, geothermal energy, photovoltaic energy, solar thermal power plants and storage technology, biogas, bio fuels and solid biomass. Renewable energies can create income for farmers operating wind and solar parks, can achieve a reduction of costs in the agricultural sector, might tackle climate change, and can achieve energy security. Decentralized solutions of energy supply encourage farmers and agricultural holdings to cooperate with local networks, power stations, and locally operating management companies for additional income and cost reduction.

Digital Technologies from Urbanized Areas Used in Rural Areas

Agriculture, rural development, and urbanization are interconnected and build up the driving force for technological progress, as the German Minister for Cooperation and Development, Mueller (2016) has pointed out at the UN Habitat conference in Berlin. Digital startups from universities, research institutes, and R&D departments of large industrial enterprises located in cities have initiated the digital revolution in Europe. Farmers benefit from large-scale production by the manufacturing industries. The digitalization of the whole supply chain including the logistics sector allow a more efficient distribution of agricultural products and food from rural areas, processed in urbanized areas to the consumers living in the cities.

Agriculture 4.0 and Bioeconomy Provide Tools for Small Holders in Developing Countries

The concepts Agriculture 4.0 and Bioeconomy might be useful also in developing countries to overcome malnutrition, to achieve the Sustainable Development Goals of the United Nations and to overcome the expected decrease of arable land due to the increase of urbanization. Digital innovations for agricultural machinery and facilities, Bioeconomy and the use of Biomass and renewable energies in agricultural production, crops plantation, live stock farming, marketing, and sales set up basic elements of the new Common Agricultural Policy of the EU (EU Commission, 2014 a, b).

Large-scale farms in developing countries can already profit from the use of modern high-tech machinery, ICT-equipment, the Geographical Information System (GIS) and earn income by large-scale wind and solar parks and Biomass power stations. The poor financial situation of family farms, small-holdings, and subsistence farming needs international support to make use of the Agriculture 4.0 and Bioeconomy strategies. High costs for implementing Agriculture 4.0 technologies enforce small holders to set up cooperatives and to set up concepts of sharing machinery, equipment, and premises and putting together plots of land. Cooperatives are a promising legal frame for small holders to get access to financial means and to profit from Bioeconomy, renewable energies, and the achievements of Agriculture 4.0. Inefficient subsistence farming and civil war problems are threats for a successful agricultural policy in developing countries. The expected decrease of arable land and forestry by urbanization is a further threat on future growth and for tackling the pace of population increase (FAO, 2015, G 7 Summit, 2016).

Fragmented Digital and Energetic Infrastructure and Financial Strains on Agricultural Holdings

The use of high-tech-systems, ICT-systems, and of digitally equipped machinery in agricultural production needs high speed 5 G digital networks and a functioning energetic infrastructure in all countries, worldwide. Especially rural areas in Germany and Europe face a fragmented digital and energetic infrastructure as crucial burden to the bio economic transformation of traditional agricultural production to smart farming. The gap between increasing prices of machinery, equipment, and services and decreasing prices for food sales is a financial strain for the agricultural sector. Thus, high investment costs and financial risks for implementing the Agriculture 4.0 and Bioeconomy concepts are a crucial strain on agriculture (Torrìsi, G. 2009). Strains on the agricultural sector by the digitization of agricultural production and the use of the internet are also to be feared by cyber crime and by the hacking of digital systems and data banks.

Impact of Agriculture 4.0 and Bioeconomy

Official data and results on the impact of the new agricultural policy, the multifold official programs and of public procurement activities on the economy and the agricultural sector are poor, as official statistics do not adequately cover the basic aspects of Bioeconomy and Agriculture 4.0 covering the risks, financial strains and benefits from the transformation. The state of the transformation from traditional farming to digitalized agricultural production and to the use of Bioeconomy and renewable energies is not yet transparent. A few reports on survey results conducted by consulting groups and company associations approve that about 10-15% of SMEs in manufacturing industries in Germany have implemented the concept Industry 4.0, whereas the majority of farmers in Europe are at the very beginning of Rifkin's digital revolution (Boston Consulting Group, 2015). The poor official database is also criticized by the Thuenen Institute (2018), which has set up a Bioeconomy Monitoring System. The institute admits: "Even the question of the actual state of bio-economy in Germany could not satisfactorily be answered, as detailed data do often not exist."

SOLUTIONS AND RECOMMENDATIONS

The use of High-tech-systems and ICT-systems in agricultural production and food processing could enable Millions of farmers and manager of holdings and the nutrition industries to save of resources and input, for overcoming transformation costs and risks. Especially for small holders and for family farms the sharing of High tech agricultural machinery and equipment and the founding of cooperatives might strengthen their financial capacities and support to tackle financial strains and risks of investment in implementing digital and bio-based innovations. Enterprises of the machine manufacturing industries can save resources by improving recycling procedures. This will significantly reduce costs of agricultural production and the use of innovative machinery. The public sector is to invest more in digital and energetic infrastructures especially in rural areas. More public and private funding is necessary, and the release of "green loans" and "green equity funds" and temporary tax reliefs might lead to solutions.

The official database is still inadequate. To analyze the impact of the new strategies and to draft comprehensive progress reports the official database is to be improved. Governments are proposed to provide more official macro- and microeconomic information about the benefits and risks of the transformation from traditional farming to smart and digitized farming. Statistical institutes and the agricultural ministries are to improve their standard data collections. Governments are to increase expenditures for restructuring the fragmented digital and energetic infrastructures especially in rural areas for the benefit of the agricultural sector.

FUTURE REASEARCH DIRECTIONS

The world is staying at the beginning of the digital revolution. More research is needed on the micro- and macro-economic level. For comprehensive and accurate analyses and reports on the impact of the implementation of the strategies Agriculture 4.0, Bioeconomy, including renewable energies, in the agricultural sector accurate official databases are a precondition. Agricultural Ministries and Statistical Institutes are to set up accurate databases to cover the transformation from the traditional economy to a digitized agricultural sector. Official data have to cover the benefits and the costs of investment for digital

and bio-based innovations, business results on production, income, profitability, risks, and employment, and the financial needs of agricultural holdings and family farms for investment in digitization.

CONCLUSION

The new strategies Agriculture 4.0 and Bioeconomy of the German Government and the EU Commission are promotion programs to foster digital technologies and bio-based innovations by funding R&D. The discussed programs provide also opportunities for farmers and agricultural holdings to make use of digital technologies, high tech machinery, bio-based innovations, and renewable energies.

Millions of agricultural holdings in Europe are encouraged to use biomass, renewable energies, digital innovations, and high tech machinery provided by the industrial and digital sectors under the label of Industrie 4.0. The officially funded programs, platforms, fairs, and conferences aim to achieve more efficiency and sustainable growth in the agricultural sector for securing self-sufficiency in food and energy production in Europe and to cover the Sustainable Development Goals of the United Nations. Both concepts may be useful to developing countries to win the race against exploding population and the sharp increase of demand for food.

Threats on implementing the digitization in the agricultural sector are to be feared by the risks of high investment costs, cyber crime, and by poor public budgets for improving the fragmented digital and energetic infrastructures.

The world is staying at the beginning of the digital revolution. Up to date these digital innovations have not yet been implemented by the majority of companies and holdings in the industrial and agricultural sectors. Due to poor official statistics on micro- and macro-level, only inaccurate reports on the impact of the digital and bio-based technologies of Agriculture 4.0 and Bioeconomy on holdings in the agricultural sector are available. Official databases covering of the risks, benefits, growth, and competitiveness during the digital transformation need significant improvements.

REFERENCES

AGCO. (2019). *Manufacturing*. retrieved in April 2019 from <https://www.agcocorp.com/about/manufacturing.html>

Agritechnica. (2017). *Systems & Components at Agritechnica is all about connections*. International Fair for Agricultural Machinery. Retrieved in May 2018 from http://www.dlg.org/archive_agriculture.html?detail/2015dlg/1/2/9185

Bailey, P. (2013). *Remote-controlled helicopter tested for use in vineyard applications*. Retrieved in April 2016 from website <https://www.ucdavis.edu/news/remote-controlled-helicopter-tested-use-vineyard-applications-video>

Boston Consulting Group. (2015). *Crop Farming 2030, The Reinvention of the Sector*. Author.

Commission European Union (E.U.). (2014a). *Where next for the European Bioeconomy?* Retrieved in March 2015 from <http://ec.europa.eu/research/bioeconomy>

Commission E. U. (2014b). *Agriculture, The EU's Common Agricultural Policy (CAP): for our food, for our countryside, for our environment*. Retrieved in April 2016 from website http://europa.eu/pol/index_en.htm

Commission E. U. (2017). *Directorate-General for Research and Innovation: Bioeconomy strategy review, A Commission staff working document*. Retrieved in July 2018 from <https://publications.europa.eu/en/publication-detail/-/publication/1f0d8515-8dc0-4435-ba53-9570e47dbd51>

Commission E. U. (2018a). *The European Bioeconomy in 2030, Delivering Sustainable Growth by addressing the Grand Societal Challenges*. Retrieved in May 2018 from <http://www.epsoweb.org/file/560>

Commission E. U. (2018b). *Agricultural technological innovation*. Retrieved in April 2019 from <https://ec.europa.eu/jrc/en/research-topic/agricultural-technologicalinnovation>

Commission E. U. (2018c). *White Paper on the European Bioeconomy 2030*. Retrieved in April 2019 from website http://www.greengrowthknowledge.org/sites/default/files/downloads/resource/BECOTEPS_European%20Bioeconomy%20in%202030.pdf

EU Commission's Science and Knowledge Service. (2016). *Annual-report-2016*. Retrieved in May 2018 from <https://ec.europa.eu/jrc/en/publication/annual-reports/jrc-annual-report-2016>

Commission E. U. (2018d). *Bio-based products, from idea to market, 15 EU success stories, study publication metadata*. Retrieved in April 2019 from <https://publications.europa.eu/en/publication-detail/-/publication/23ab58e0-3011-11e9-8d04-01aa75ed71a1/language-en/format-PDF/source-88043987>

FAO. (2017). *The State of Food Security in the World 2017*. Retrieved in April 2019 from website <http://www.fao.org/3/a-i7695e.pdf>

FAO. (2015). *The State of Food Insecurity in the World 2015*. Retrieved in April 2019 from <http://www.fao.org/publications/sofi/en/march>

Fraunhofer Institute Germany. (2018). *What does the Bioeconomy of the future look like? Press Release of 10.1.2018*. Retrieved on May 25, 2018 from <https://www.isi.fraunhofer.de/en/presse/2018/presseinfo-01-2018-biokompass.html>

G7 Summit. (2016). Retrieved in June 2016 from website <http://www.japan.go.jp/g7/summit/>

German Agricultural Association. (2015). *Press releases and Conference papers of DLG*. Retrievable from website http://www.dlg.org/homeen.Html?&no_cache=1&L=51

German Agricultural Association. (2017). *Systems & Components at Agritechnica is all about connections*. Retrieved in May 2018 from website http://www.dlg.org/archive_agriculture.html?detail/2015dlg/1/2/9185

German Agricultural Association. (2018). *DLG Competence Center, Solutions for the future, internet platform*. Retrieved in April 2018 from http://www.dlg.org/competence_center_agriculture.html

German Bioeconomy Council. (2018). *Bioeconomy Policy (Part III) Update Report of National Strategies around the World*. Retrieved in May 2018 from website http://biooekonomierat.de/fileadmin/Publikationen/berichte/GBS_2018_Bioeconomy-Strategies-around-the_World_Part-III.pdf

German Energy Agency. (2016). *Renewables - made in Germany 2016*. Retrieved in April 2019 from https://www.dena.de/fileadmin/dena/Dokumente/Pdf/9079_Flyer_Anbieterkatalog_renewables_Made_in_Germany_2016.pdf

German Farmers' Association. (2015). *Conference papers*. Retrieval from website <http://www.bauernverband.de/information-english>

German Research Center Juelich. (2018). *Sustainable Bioeconomy, Plant research*. Retrieved on May 25, 2018 from http://www.fz-juelich.de/portal/EN/Research/bioeconomy/_node.html

International Fair on Industries Hanover. (2018). *Press statements, conference papers*. Retrieved in April 2018 from <http://www.hannovermesse.de/en/info/for-exhibitors/trade-fairline-Up/digital-factory/>

McClelland, J. (2015). *Top 5 tech innovations in agriculture publ*. Retrieved in April 2019 from website <http://raconteur.net/sustainability/top-5-tech-innovationsin->

Minister for Cooperation and Development. (2016). *German UN-Habitat Forum in Berlin 2016*. Retrieval from website <http://www.german-habitat-forum.de/>

Minister of Food and Agriculture. (2015). *International Agriculture Ministers' Summit in Berlin. Press release no. 24 /2015*. Retrieved from website <http://www.bmel.de/SharedDocs/Pressemitteilungen/EN/2015/024-SC-GFFA>

Minister of Food and Agriculture. (2019). *The decision to establish an international Digital Council is a breakthrough!* Retrieved in April 2019 from <https://www.bmel.de/SharedDocs/Pressemitteilungen/EN/2019/abschluss-gffa-2019.html>

Ministry for Economic Affairs and Energy. (2015). *Joint Industry 4.0 platform launched*. Retrieved in April 2015 from <http://www.bmwi.de/en/press>

Ministry of Education and Research. (2011). *National Research Strategy BioEconomy 2030*. Retrieved in 2015 from website <http://www.bmbf.de>

Ministry of Education and Research. (2014). *Bioeconomy international 2014*. Retrieved in May 2015 from website http://www.bmbf.de/pub/bioeconomy_2030.pdf

Ministry of Food and Agriculture. (2015). *EU agricultural policy*. Retrieved in May 2016 from http://www.bmel.de/EN/Agriculture/EU-AgriculturalPolicy/EU-agriculturalpolicy_node.html

OECD. (2015). *The Bioeconomy to 2030, Designing a policy agenda*. Retrieved in April 2019 from <http://www.oecd.org/futures/bioeconomy/2030>

OECD-FAO. (2015). *Agricultural Outlook 2015-2024*. Retrieved in April 2019 from <http://www.oecd.org/>

Patermann, C., & Aguilar, A. (2018). The origins of the Bioeconomy in the European Union. *Journal New Biotechnology*, 40, 20–24. Retrieved in May 2018 from Website www.elsevier.com/locate/nbt

Price Waterhouse. (2015). *Strategy & Industry 4.0, Opportunities and challenges of the Industrial internet January 2015*. Retrieved from <http://www.strategyand.pwc.com/media/file/Industry>

Rifkin, J. (2015). *The Industrial Revolution*. Retrieved in March 2015 from <http://thethirdindustrial-revolution.com/>

Rossner, E. (2015). Der Dritte Weg, Interview with Jeremy Rifkin in German. *Innovations*, 33/34. Retrievable from www.innovationsmanager-magazin.de

The Thuenen Institute. (2018). *Bio-economy Monitoring*. Retrieved in May 2018 from <https://www.thuenen.de/en/institutsuebergreifende-projekte/bio-economy-monitoring/>

Torrise, G. (2009). *Public infrastructure: definition, classification, and measurement issues*. Retrieved in May 2015 from http://mpira.ub.unimuenchen.de/12990/1/survey_infra_def.pdf

United Nations. (2015). *Transforming our World, the 2030 Agenda for sustainable Development*. Retrieved in April 2019 from website <http://sustainabledevelopment.un.org>

Van Vark, C. (2016). Drones set to give global farming a makeover. *The Guardian*. Retrieved from <http://www.theguardian.com/globaldevelopment/2015/dec/26/drones-farming-crop-problems-uavs>

ADDITIONAL READING

Bugge, M., Hansen, T., & Klitkou, A. (2016) What Is the Bioeconomy? A Review of the Literature, retrieved on April 29, 2019 from <https://www.mdpi.com/2071-1050/8/7/691>

Mechanical Engineering Federation Germany. (2019) Rechercheportal - Forum Industrie 4.0 -VDMA, retrieved on April 29, 2019 from website <https://industrie40.vdma.org/en/rechercheportal?>

The Federal Government (Germany). National Bioeconomy Strategy, Summary, retrieved July, 2020 from https://www.bmbf.de/files/2020_1501_National-Bioeconomy-Strategy_Summary_accessible.pdf

KEY TERMS AND DEFINITIONS

Agriculture 4.0: A strategic concept to implement digital technologies and high-tech machineries and equipment in the agricultural sector and in the nutrition industries.

Bio Energy: Is set up by biogas, bio-fuel, etc.

Bio Mass: Encompasses any biological material from agriculture and forestry and includes animal-based material as a product in itself or to be used as raw material for renewable energy such as bio gas or even 3 D printing.

Bioeconomy: An official strategy to implement a new economic system based on the sustainable use of biological resources and is including the traditional concept of “biotechnology.”

Precision Farming: The use of digital equipment enabling farmers for a more efficient and precise use of seeds, fertilizer, and pesticides for crop plantation.

Renewable Energies: Are defined by solar, photovoltaic, wind, bio energy, etc.

Overview of Solid Waste Management of Healthcare and Related Organisations

Isaiah Adesola Oke

 <https://orcid.org/0000-0002-7082-7682>

Obafemi Awolowo University, Nigeria

Lukman Salihu

University of Hafr Al-Batin, Saudi Arabia

Idi Dansuleiman Mohammed

Land, Air, and Water Consulting, Nigeria

Asani M. Afolabi

Ladoke Akintola University, Nigeria

INTRODUCTION

Medical solid waste is any solid waste which is generated in the diagnosis, treatment (provision of medical services), or immunization of human beings or animals, in research pertaining thereto, or in the production or testing of biological. Medical waste (MW) is classified by the World Health Organization (WHO) as “waste that is generated in the diagnosis, treatment or immunization of human beings or animals”. As defined by WHO (2014), the term health-care waste includes all the waste generated within healthcare facilities, research centres and laboratories related to medical procedures. It also includes waste produced in the course of health care undertaken in the home. Management of medical (healthcare) waste is an integral part of infection control and hygiene programs in healthcare settings. These settings are a major contributor to community acquired infection, as they produce large amounts of biomedical waste. Biomedical waste can be categorized based on the risk of causing injury and/or infection during handling and disposal. Wastes targeted for precautions during handling and disposal include sharps (needles or scalpel blades), pathological wastes (anatomical body parts, microbiology cultures and blood samples) and infectious wastes (items contaminated with body fluids and discharges such as dressing, catheters and IV. lines). Other wastes generated in healthcare settings include radioactive wastes, mercury containing instruments and polyvinyl chloride (PVC) plastics. These are among the most environmentally sensitive by-products of healthcare (Remy, 2001). World Health Organization (WHO) stated that 85% of hospital wastes are actually non-hazardous, around 10% are infectious and around 5% are non-infectious but hazardous wastes. In the USA, about 15% of hospital waste is regulated as infectious waste. In India this could range from 15% to 35% depending on the total amount of waste generated (Glenn and Garwal, 1999; Soliman and Ahmed, 2007).

Healthcare waste products should be considered as a reservoir of pathogenic microorganisms, which can cause contamination and give rise to infection. If waste is inadequately managed, these microorganisms can be transmitted by direct contact, in the air or by a variety of vectors, and can pose a serious threat to human health and to the environment. The inefficient handling of biomedical waste is more

DOI: 10.4018/978-1-7998-3479-3.ch091

likely to cause problems such as blood borne pathogens to the groups at highest risk, namely; healthcare staff, scavengers, and municipal workers (from needle sticks for example, if the biomedical wastes are handled and disposed together with domestic wastes). In many countries, cities are with inadequate segregation of hazardous and non-hazardous biomedical wastes, as well as a lack of suitable waste treatment facilities and methods. These cities also have either unregulated or non-existent legislation, with regard to waste processing and treatment, in addition to inefficient training of personnel and lack of personal protective measures (Askarian *et al.*, 2004; Marinkovic *et al.*, 2005; Rasheed *et al.*, 2005; Soliman and Ahmed, 2007).

Medical solid wastes management is an important factor in environmental hygiene and needs to be integrated with total environmental planning (Oke, 2008). Elements of waste management include characterisation, storage, collection, transportation, treatment and final disposal. Although, the World Health Organization (WHO, 2004), US Environmental Protection Agency (EPA, 1991), US Centres for Disease Control and Prevention, Italy (Liberti *et al.*, 1994, 1996), Finland, Japan and Germany (Miyazaki *et al.*, 2007; Miyazaki and Une, 2005; Tsakona *et al.*, 2007) have already established strict guidelines for the management of infectious waste materials, solid waste management is a major problem in most developing countries of the world due to its ever growing and endless generation coupled with poor management (Longe and Williams, 2006, Meghdad *et al.*, 2016).

BACKGROUND

In some countries, solid and medical wastes disposal is governed by laws of the Ministry of Environmental Affairs in collaboration with the Ministry of Health and Population. Regulations classify the waste from healthcare settings to be hazardous, and state that certain precautions should be considered during collection, handling and final disposal (Soliman and Ahmed, 2007, Meghdad *et al.*, 2016). One of the most important and initial steps in the development of a plan or in the performance of risk or cost analyses in the field of solid waste management involves a thorough understanding of the quantities and properties of the material that needs to be evaluated or treated (Diaz *et al.*, 2008). In this particular case, the material under consideration is healthcare waste. Unfortunately, there is a limited quantity of reliable information in the open literature on the quantities and characteristics of the various types of wastes that are generated in healthcare facilities (Diaz *et al.*, 2008). Furthermore, most of the reports dealing with analyses and evaluations conducted on the performance of a particular type of treatment method either in industrialized or in developing countries do not clearly specify or characterize the quality of the waste undergoing the treatment. Article provides basic information on the quantities, characteristics; collection, transportation and disposal of the solid wastes generated in various types of healthcare facilities located in industrialized and in developing countries are rare (Diaz *et al.*, 2008). This therefore calls for documentations on such important waste with a particular attention to prevention spreading of infectious diseases and to aid environmental engineer in the design of disposal facilities. The main objective of this study is to presents documentations on health care solid wastes with a particular attention to prevention spreading of infectious diseases and to aid environmental engineer in the design of disposal facilities.

Medical Wastes Treatment and Disposal Practices

Treatment of toxic and infectious waste has been defined as any method, technique or process designed to change the biological character or composition of waste to render it non-toxic or non-infectious. There

are several factors to consider when selecting among the medical waste treatment methods (Manyele *et al.*, 2003, PATH, 2005, Visvanathan, 2006, Paweł, 2019). In the selection of treatment process for medical waste the capital cost, the operating costs, the running cost of pollution control and waste preparation equipment, operating charges, sterilization efficacy, maintenance and higher operator skills, air emissions, water emissions and the characteristics of treated waste are factors that must be considered (Manyele *et al.*, 2003).

Waste Collection, Transportation, Storage and Segregation (sorting)

Collection and transportation of medical waste remains a serious concern in most of the countries in the world as improper collection and transportation endanger the environment.. Collect and store compatible wastes in strong, tight containers in a secured area (protected from the weather, such that none can escape by gravity into the environment. Keep lids tightly secured when not in use. Medical waste transportation refers to the haulage and handling of waste from inside healthcare facilities to treatment sites, which can either exist on-site at a hospital or be a central off-site facility. A second transportation phase typically occurs when the treated waste residual, typically ash from an incinerator or waste sterilized through autoclaving or microwaving, is moved to a landfill for final disposal

Segregation of waste in a hospital affects infection control, patient safety, occupational health (in the hospital), and overall community health and the environment. It also helps to cut down emissions of greenhouse gases, persistent organic pollutants (POPs), mercury and other hazardous waste. At hospitals and other healthcare facilities, waste is generally sorted into color-coded bins or bags, with each receptacle denoting a different waste stream or waste type. The colour selected for each waste type, along with what types of waste go into each stream, varies from region to region, with some using the source of the waste as a basis for sorting, while others use the likelihood of objects (pathogenicity to determine its disposal waste stream).

Medical wastes are stored in containers with colour code (Visvanathan, 2006, Paweł, 2019). Colour is used to differentiate hospital waste at the generation point. Infectious wastes are stored in a yellow marked, strong leak proof bag or container. Chemical and pharmaceutical wastes are stored in a brown marked plastic bag or container. Black marked plastic bags in containers are used for storing general waste. Radioactive wastes are stored in a red lead box labelled with a radioactive symbol. Sharps are stored in a yellow marked puncture proof container with covers on them. Infectious, pathological and chemical wastes are usually collected in red plastic bags (Nemathaga *et al.*, 2008). In Serbia segregated pharmaceutical waste is packed in the red containers or red bags, while cytostatic waste is packed in purple containers or bags. Infectious waste is packed in yellow bags or containers; and other hazardous waste (chemical) - in red containers. In the most EU countries there is the same colour coding system since Serbian model for colour coding of healthcare waste streams is taken from the EU countries experience and practices.

Medical Wastes Treatment Methods: Methods of infectious wastes treatment are well documented in literature, but its applications varied from countries. The general or common methods are incineration, burying (landfills) the wastes, open burning, autoclave, chemical treatment, hydropulping, and compaction.

- **Open Burning:** If infectious and medical wastes cannot be taken to an incinerator, they may be burned in a drum, protected hearth, or an open pit. Drum burning is preferred to open pit or open burnings because combustion is better and temperatures are higher. The advantages of burning in

an open drum are that its materials are available locally and it has a low cost. The disadvantages include the followings

- a. Requires supervision for proper usage.
- b. Does not burn as cleanly as a special incinerator .there may be smoke and environmental emission hazards.
- c. Fuel is required.
- d. Needles will not be destroyed, so residual metal and ash remain a hazard (though they are not infectious after burning).
- e. Hot, molten plastic is a serious hazard and burning pits may not work well in the wet season.

An open oil drum is among the best solution for burning. The drum, with top and base removed, are placed on a steel grid, which is raised off the ground with bricks to allow good air circulation. If a drum is not available, an open brick hearth can be used instead. Sweep molten plastic and ash into an ash pit behind the hearth. Finally, if neither burning in drum nor in concrete drum is feasible, burn waste in an open pit. Take care to ensure that all the infectious wastes are burned because shallow pit burning is not very efficient. Burning infectious and medical wastes in an open drum, protected hearth or pit is the next environmental friendly method to incineration. It should be noted that burning waste should be protected so that wind does not blow hot ash and so that molten plastic cannot burn people near the fire.

- **Burying and Land Fill:** If for some reasons infectious and medical wastes cannot be burned, the wastes can be buried in a pit or in an approved land fill site. Literature such as Askarian *et al.* (2004 a and b); Marinkovic *et al.* (2005 and 2008); Oke (2008); Rasheed *et al.* (2005); Daiz *et al.* (2005), Batiu *et al.* (2018) present facilities for burying infectious and medical wastes.

The advantages of burying infectious and medical wastes are:

- a. No additional equipment required and no smoke.
- b. No fuel and maintenance are needed.
- c. the capital, operational and maintenance costs are low.

The disadvantages of the treatment method are as follows:

- a. Digging deep pits and covering them places enormous burden on staff (is a lot of work)
- b. Syringes can be unearthed accidentally (or on purpose).
- c. Pits cannot be used when the water table is near the surface because liquids leak into ground water and infectious and medical wastes may surface over time.
- d. Hard ground may make digging difficult.
- e. infectious and medical wastes are bulky and some infectious and medical wastes may never degrade
- f. The environmental cost is high and much land is needed

This option should only be used when: no incinerator is available, burning is not acceptable or possible and suitable land is available (the land should be securely protected and not accessible to children and the community, the water table should be at least 1.50 m below the base of the pit, even during the wet season and the ground must not be too hard). Burying infectious and medical wastes in the ground is the least safe method because of its tendency to pollute ground and surface water. Since landfill operations may cause loss of containment integrity and dispersal of infectious waste, it is recommended that all infectious waste be treated prior to disposal. Open pit burning or burial of medical waste should be discouraged. To control and minimize the levels of environmental pollution and occupation health injuries from incineration processes, the hospital must control the

waste stream, for example, avoiding certain materials in the waste stream and sorting the wastes at the point and time of generation.

- **Incineration:** Incinerator in developing countries is self-made and is constructed from burned bricks and cement; it has a shape of house fireplace. Waste is burned using coal as fuel, which does not allow proper control of temperature. A high amount of ash is generated because of the incomplete burning of waste. Its chimney is also short and, depending on wind direction, emitted gases are dispersed to nearby communities, causing nuisance and being a potential cause of bronchitis and pulmonary ailments such as asthma (Mato and Kassenga, 1997 Mato *et al.*, 1994). Waste that is improperly incinerated, especially that containing plastic materials is known to give rise to toxic gases such as dioxins and furans that are carcinogenic (Lee *et al.*, 1996). Plastic materials as part of the waste encourage the release of dioxins when incinerating them. This is because plastics contain chlorine by-products such as polyvinyl chlorides (PVCs, Mato and Kassenga, 1997). The problems of medical waste treatment are typical of any developing country that cannot afford to buy more environmentally friendly incinerators with the latest technology (Nemathaga *et al.*, 2008). Visvanathan (2006) reported that incineration is the common methods in Asian countries. Figures 1 and 2 show typical incinerators fabricated using local materials in Nigeria. Incineration is a perfect method for all types of HCW, both hazardous and non-hazardous (Hasan and Rahman, 2018, Abu Mhady *et al.*, 2019).
- **Autoclaving:** The autoclaves have a temperature range of 50°C –250°C, but they are operated at 160°C as the optimum temperature to kill bacteria and infectious diseases, but not to melt any materials that are introduced. Autoclaving is now recognized as an important method for treating various kinds of infectious hospital waste. Its major drawback is that it cannot handle large quantities of hazardous waste. Sharps are generally known to be capable of transmitting diseases. Safe handling and disposal of sharps is an essential part of any infection control program (Nemathaga *et al.*, 2008). Autoclaving is suitable for sterilizing hazardous waste. However, the volume of waste cannot be reduced in this process. It is not suitable for recognized body parts (Hasan and Rahman, 2018). On the other hand, the hydropulping, an oxidation technique, has the required water content of approximately 80% as a result of an increase in weight; however, the resultant volume can be as little as 30% of the original volume. The use of a pulping system is highly controversial for clinical waste treatment. The compaction system can reduce the waste volume up to 60%. However, sterilization is not possible in this method. A thorough description of technology with their strengths and weaknesses, as well as some key factors for technology selection, has been presented in preceding publications (Hasan and Rahman, 2018).
- **Open dumping:** Open dumping has long been recognized as a potential source of public health and environmental problems. Generally landfill has accepted as the final option of resting hazardous hospital solid waste. Open dumping, because of its inherent problems such as leakage of toxic substances into the environment; easy access from insects, rodents and other small animals, most of which are disease vectors, has been replaced by landfilling in the management of solid wastes (Nemathaga *et al.*, 2008).
- **Landfill:** The main idea of the landfill is for storage and containment of the waste deposited into it. At present, the landfills are operated like an open dump in some countries. Each day, the general and medical wastes are dumped in the landfill and later burned. Burning is aimed at reducing the volume of waste and stopping the spread of papers. This process is repeated until the landfill is full, and then the landfill is covered with soil. In normal operation of the landfill, the wastes are spread over, compacted and then covered with a layer of soil on a daily basis. This eliminates the

need for burning the waste as practised some hospitals. The burning itself is a potential source of generating toxic chemicals. This is more likely since wastes such as plastics, syringes and paper are burned together. There is high chance that toxic chemicals like dioxins and furans are generated. The community living near the landfill therefore acts as passive samplers through inhalation of polluted air. The landfills at the hospitals are supposed to be divided into two portions, one for putting general waste and the other, medical waste. Only general waste is supposed to be burned. However, in practice this is not followed. Both general and medical wastes are mixed at the landfill and burned together (Nemathaga *et al.*, 2008).

- **Recycling:** Recycling of solid waste at some hospitals in developing countries is being undertaken at a small scale level. Most the waste that is being recycled is leftover food. Other general wastes that have potential for recycling but that are not recycled include plastic bags, plastic containers and cans. These plastics come from various sources within the hospital, but the major percentage comes from cafeteria related service (25%) and medical packaging (26%, Lee *et al.*, 2002). Most of the cans also come from cafeteria related services and within the hospitals; these were mainly empty cans of drinks or beverages used by patients, staff and visitors at the hospital. Plastics and cans have a high potential to be recycled (Nemathaga *et al.*, 2008).
- **Disinfection:** In order to reduce the toxicity of some medical wastes, chemical disinfectants (chlorine dioxide, sodium hypochlorite, or per acetic acid) are sometimes used. For solid wastes, disinfection is effective if only waste materials are shredded. In some cases, the disinfectants themselves are hazardous, thus it is not recommended for treating pharmaceutical, chemical and some types of infectious waste (Zarook, 2012).
- **Disinfection by Plasma:** In this process, low temperature plasma which is produced by the plasma generator using air as working fluid organizes a combustion process. The medical waste is constantly mixed, thus it maximizes the heat and mass exchange which saves any energy loss. The heat produced is used as an additional heat source in the process. This technology eliminates the formation and release of irregular forms of NO_x and high-toxic substances (dioxins) into the atmosphere. Another main advantage is that it has low consumption of energy compared to other mineralization (combustion) processes (Zarook, 2012).
- **Emerging Technology:** A new technology for management of hazardous medical waste that transforms the regulated medical waste into municipal solid waste is recently introduced. This method involves shredding and grinding the infectious medical waste bags via sharp cutting blades that are installed within the vessels. The blades rotate around 1750 revolutions per minute and the volume of the shredded waste is reduced by 80%. The steps included in the process are loading, shredding, heating, sterilization, cooling, draining, vacuum and unloading. The whole process is enclosed in a compact system and there is no intermediate handling of the waste within the process. Due to the compact size, this system can easily be used for on-site treatment of the waste and installed in hospitals. This will reduce the transportation costs of the medical wastes. In terms of environmental aspects, it is a clean and chemical-free technology and does not have any hazardous emission or radiation. This method is economical and environmentally friendly and is reliable in terms of ease of use and maintenance. This technology is currently practiced in the middle-eastern countries such as Iraq, Jordan, Kuwait, Lebanon, Syria and United Arab Emirate (Zarook, 2012).

Medical Wastes Management Techniques

Improved health-care waste management requires awareness raising; a legal framework and a policy, which have to be implemented locally; waste management plans for health-care establishments; staff training. The elements of a well-documented infectious waste management plan can be outlined as designation and identification, segregation, packaging, storage, transport and handling, treatment techniques, disposal of treated waste, contingency planning and continued staff training (Manyele, 2004; Visvanathan (2006), Blackman, 1996; Griffin, 1989). The step by step waste management techniques are generation, segregation (sorting) storage, transportation, treatment and final disposal.

- Waste generation in both developed and developing Nations:** Bakiu and Durmishaj (2018) reported that highly developed countries have a larger production of medical waste than middle developed and developing countries. Highly developed countries produce 1.1–1.2 kg per capita, 0.4–0.5 kg of which is hazardous waste; middle developed countries produce 0.8–6 kg per capita, 0.3–0.4 of which is hazardous waste; and developing countries produce 0.5–3 kg of waste per capita. In Croatia annual waste production per capita is 1.2 kg, out of which 0.16 kg is hazardous medical waste. Comparison of world data shows large differences in daily medical waste production between affluent and poor regions. North America produces 7–10 kg of waste per hospital bed daily, Western Europe 3–6 kg, whereas South America produces 3 kg, and Eastern Europe 1.4–2 kg per bed. The difference in quantities results from the fact that developed countries invest much more money in health systems, leading to larger amounts of medical waste generation. Meghdad *et al.*, (2016) stated that medical waste production rate per active beds in selected hospitals in Iran were comparatively less than Tehran-based (2.71 kg/day.bed) and Zanjan-based (2.402 kg/day. bed) hospitals. Visvanathan (2006) reported that medical waste generation in some countries as 0.8–1.67 kg/bed/day in Bangladesh, 0.27 kg/bed/day in Bhutan, 1.0–2.0 kg/bed/day in India, 1.90 kg/bed/day in Malaysia, 0.5 kg/bed/day in Nepal, 1.06 kg/bed/day in Pakistan, 0.36 kg/bed/day in Sri Lanka, 0.68 kg/bed/day in Thailand and 2.27 kg/bed/day in Vietnam. Paweł (2019) stated that medical waste generation in European's countries as 1.22, 1.76, 0.13, 0.63, 0.40, 1.39, 1.23, 0.23, 1.64, 4.80, 0.06, 0.01, 1.46, 5.03, 2.40, 0.20, 0.19, 5.07, 0.57, 0.57, 0.76, 3.79, 0.32, 4.12, 2.68, 2.40, 0.63 and 22.52 kg/bed/day in Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and United Kingdom respectively. Hasan and Rahman (2018) stated that healthcare waste generation rate and hazardous healthcare waste generation rate in Khulna city, Bangladesh were 0.90 kg bed⁻¹ day⁻¹ and 0.18 kg bed⁻¹ day⁻¹, respectively. Abu Mhady *et al.* (2019) documented that medical waste in Greece was varied from 0.00124 kg/bed/day (private psychiatric hospitals) to 0.718 kg/bed/day (public university hospitals), whereas in Bangladesh, the rate was 0.57 kg/patient/day. The rate of medical waste in Palestine was between 0.33 and 0.84 kg/patient/day, but updated report found 1.306 kg/bed/day and an occupancy rate of 80% other reports found 0.85 kg/bed/day. Nevertheless, the generated waste in Iran included 51.6% of infectious waste, 47.2% general waste and 1.2% sharps. Minas *et al.* (2017) presented medical waste generation from all the regions as follows:

- In Africa medical waste generation (kg/ bed/ day) for Algeria (0.96), Cameroon (0.55), Egypt (1.03), Ethiopia (1.1), Mauritius (0.44), Morocco (0.53), Sudan (0.87), Nigeria (1.68) and Tanzania (0.75);

- b. In Asia medical waste generation (kg/ bed/ day) for Bangladesh (1.24), Malaysia (1.9), China (4.03), Pakistan (2.07), India (1.55), Indonesia (0.17), Palestine (2.02), Indonesia (0.75), Thailand (2.05), Iran (3.04), Turkey (4.55), Japan (2.15), Nepal (0.5), Jordan (2.69), Lebanon (5.7), Korea (2.4), Kazakhstan (5.34), Laos (0.51) and Vietnam (1.57);
 - c. In America medical waste generation (kg/ bed/ day) for Argentina (3.0), Ecuador (2.09), Brazil (2.94), El Salvador (1.85), Canada (8.2) and USA (8.4); and
 - d. In Europe medical waste generation (kg/ bed/ day) for Bulgaria (2.0), Netherlands (1.7), Italy (4.0), Norway (3.9), France (3.3), Spain (4.4), Germany (3.6), Latvia (1.18), Greece (3.6) and UK (3.3).
- **Composition of medical wastes:** One important component of a waste characterization program involves the determination of the composition of the wastes. A sound understanding of the contents of the waste stream, in particular the contents of the non-infectious fraction of the waste is helpful in the development of realistic waste reduction and recycling programs. The composition of hospital wastes generated in institutions in Peru shows that a substantial amount of the wastes consists of paper products, plastic and textiles in the form of cotton and gauze. This medical solid waste generation depends on several factors such as institutions mode of operation, unit, number of patient, number of beds and rooms, habit and culture of the patient and climatic conditions.
 - **Bulk density of medical wastes:** Bulk density is defined as the uncompacted mass of a material occupying a known volume. Bulk density generally is reported as mass per unit of volume (kg/m^3). In waste management, it is important to know the bulk density of the waste or components of the waste for a number of purposes, which include: determination of storage space, definition of size for the collection vehicle and estimation of the requirements for processing equipment (compaction, size reduction, disinfection and others). Diaz *et al.* (2008) shows that the bulk density for the entire waste stream ranged from 137 to 359 kg/m^3 , with an average of 218 kg/m^3 in Peru. The average bulk density for common (general) waste was 211 kg/m^3 , that for contaminated waste was 226 kg/m^3 and that for special waste was 139 kg/m^3 . The bulk densities of various components of healthcare wastes generated in Guayaquil, Ecuador show that the densities vary from about 56 to 960 kg/m^3 (Diaz *et al.*, 2008). The data provide the results of thermo-chemical analyses, including heating value, for several components of the waste. Other physical-chemical characteristics for hospital wastes generated in Saõ Paulo, Brazil, (by type of generator) are in the range of 56.22 to 959.71 kg/m^3 with an overall average of 261 kg/m^3 . These variations in the bulk densities can be attributed to various factors as institutions mode of operation, unit, number of patient, number of beds and rooms, habit and culture of the patient and climatic conditions. These results indicate that waste reduction techniques may not be applicable in some places.
 - **Segregation of medical wastes:** Segregation is useful since it prevents the contamination of non-hazardous waste by the hazardous waste and making the whole waste stream hazardous. The method reduces the toxicity and the volume of the waste stream. Segregation makes it easier to transport the waste. Waste is segregated depending on the quantity, composition, and the disposal method of the waste stream (Zarook, 2012). Immediately after used, infectious and medical wastes should be discarded in an appropriately colored container with a plastic liner bag. If infectious and medical wastes are to be disinfected and sent for recycling, they should not be mixed with non-infectious wastes. Despite the commendable level of segregation of immunization and medical wastes currently achieved in medical waste management practices, literature revealed that in developing countries segregated immunization wastes are sometimes mixed together by collectors at the collection points or at the dump sites on the basis of lack of provision for separate transpor-

tation or disposal sites. This indicates that the whole mixed waste could be considered as being infectious waste, which portends a serious risk to the general public and the environment (WHO, 2004; Longe and Williams, 2006)

- **Shredding of medical wastes:** infectious and medical wastes may be shredded in a hand mill or an electric shredder so as to reduce their volume. After shredding they may be buried on-site.
- **Encapsulation of medical wastes:** Encapsulation means surrounding dangerous materials with a substance that will harden. Encapsulation ensures that infectious and medical wastes are stuck in a material, so that used materials (infectious wastes) cannot be harmful or reused. Filled infectious and medical wastes containers are placed in a metal drum or a high-density plastic container up to 75% of the final capacity. An immobilizing material such as fresh cement, bituminous sand or clay is added in the container. Once dry, the container is sealed and disposed of in a land fill or buried on site.
- **Storage of medical wastes:** Infectious Waste, wherever it arises, should be disposed of as soon as it is practicable to do, so arrangements for disposal, and for storage pending disposal. Storage will differ depending upon where the waste is generated, on the nature and quantity of the waste and perhaps on geographical and/or weather constraints. In all instances, storage arrangements should:
 - a. allow safe handling of Infectious waste packaging;
 - b. allow clear distinctions to be maintained between infectious wastes intended for different disposal methods, or, in the case of storage in patients' homes, between infectious and domestic wastes, as the means of disposal are different;
 - c. allow for storage of Infectious wastes without offensive odour;
 - d. be of sufficient size to allow for delays in the onward transportation of the waste.

The local circumstances, the nature and volume of the infectious waste to be stored should be taken into account when determining storage conditions. Where necessary, advice on storage conditions should be sought from the relevant Health and Safety Adviser or Infection Control Adviser.

- **Transport of medical wastes:** Arrangements for the collection and transport of infectious waste should be:
 - a) compatible with the arrangements made for storage of the Infectious waste;
 - b) made, in the context of the Waste and Contaminated Land (FEPA, 1991) to permit ready delivery to the means of treatment or disposal which has been arranged;
 - c) be compatible (particularly in the case of Infectious Waste created in patients' homes) with the reasonable expectations of patients, health and social care and waste management personnel in relation, for example, to working practices; and
 - d) complied with health and safety and environmental legislation and good practice, especially regarding the health and safety of patients, carers, health and social care staff, waste management personnel and others.

In both developed and developing countries, all the wards have individual safety boxes, and other wastes generated had a general box of either polyethylene or other plastic. These can be regarded as internal storage facilities for waste collection and storage. In developing countries of the world, collection of immunization waste from the wards to the stations (immunization offices) is effected immediately after the EPI activities, and the frequency of collection from the immunization offices by external

waste collectors is irregular (collection could be 2 or 3 weeks after the program). In some developing countries, transportation from wards to the immunization offices is either by car, lorry, motor-bicycle or other method depending on distance and logistics. Transportation to the disposal sites is by the same methods, which are found to be irregular and not adequate.

- A. **Current Medical Waste Treatment and Disposal Practices in Developed Nations:** At hospitals and other healthcare facilities in developed nations, waste is generally sorted into colour-coded bins or bags, with each receptacle denoting a different waste stream or waste type. The colour selected for each waste type, along with what types of waste go into each stream, varies from region to region, with some using the source of the waste as a basis for sorting, while others use the likelihood of an object's pathogenicity to determine its disposal waste stream (Bakiu and Durmishaj, 2018). This lack of standardization makes effective waste sorting difficult for healthcare workers, and causes workers to err on the side of caution, disposing objects in the infectious waste stream and causing unnecessary infectious waste generation (Bakiu and Durmishaj, 2018). Another issue with medical waste disposal is ensuring that persons do not come into contact, whether accidentally or on purpose, with disposed-of infectious items. In most jurisdictions, healthcare facilities have a legal responsibility to ensure that patrons and staff do not come into contact with infectious waste once it has been placed in a disposal bin (Bakiu and Durmishaj, 2018). Studies, focused mainly on the UK, have found that hospitals do not have adequate safeguards to prevent these contacts with hazardous medical waste, and that safe-handling procedures are frequently neglected (Bakiu and Durmishaj, 2018). This inadequacy is both a source of infection and a legal liability for hospitals, should patients become ill from poor waste management practices. Further reinforcing the need for adequate safeguards in healthcare facilities, the EPA has concluded that the disease-causing potential of medical waste is greatest at the point of generation and naturally tapers off after that point. Thus, safeguarding of infectious medical waste within healthcare facilities ought to be made a top waste management priority.

Medical waste transportation refers to the haulage and handling of waste from inside healthcare facilities to treatment sites, which can either exist on-site at a hospital or be a central off-site facility. A second transportation phase typically occurs when the treated waste residual, typically ash from an incinerator or waste sterilized through autoclaving or microwaving, is moved to a landfill for final disposal (Bakiu and Durmishaj, 2018). It is common practice for healthcare facilities to have their infectious waste stream transported by a third-party firm, contracted to take the waste from the healthcare facility to an appropriate waste depot (Bakiu and Durmishaj, 2018). These firms typically collect the waste from a few central points in a healthcare facility and then transport the waste to a disposal facility that is able to safely handle medical waste. However, there are issues with the process of contracting out waste disposal. The use of third party disposal firms poses a challenge from an incentives point of view, as the waste disposal firms, or the individuals who work for them, can pocket large sums by improperly disposing of the waste.

Disposal fees for medical waste in developed countries are very high, with hospitals in the UK frequently paying in excess of £450 per tonne for contractors to dispose of their medical waste and hospitals in the United States typically paying \$790 per tonne for medical waste disposal (Bakiu and Durmishaj, 2018). These high prices create an incentive for third-party medical waste haulage firms to dispose of the medical waste without treatment in unregulated and less expensive ways, rather than transport the waste to a proper treatment facility for sterilization. In Ireland, waste truck operators can pocket over \$2000 by illegally dumping a truck full of medical waste rather than tak-

ing it to a regulated disposal site, thus creating a very strong incentive for illegal dumping (Bakiu and Durmishaj, 2018). Developed nations increasingly have to grapple with the problem of illegal medical waste dumping, which can be particularly chronic if the country has a weak infectious medical waste tracking system. Illegal dumping is a significant issue, as these untreated infectious waste deposits present a health risk to the public due to potential for pathogen release, and a drain public fund as cleanup costs for medical wastes are extremely high (Bakiu and Durmishaj, 2018). Another problem relating to illegally disposing of infectious medical waste occurs in developing countries such as India, where governments are grappling with disease outbreaks due to third-party firms receiving medical waste from healthcare facilities and then reselling items such as sharps on the black market for re-use (Bakiu and Durmishaj, 2018).

Zarook (2012) stated that a new patented technology that helps in better management and treatment (new technology for management) of the medical wastes had been developed America. Based on this technology, Med-Shred, Inc., (Texas, USA) has developed a mobile shredding and chemical disinfecting machine that is aimed for on-site treatment of hazardous medical waste (Zarook, 2012). The machine converts the medical waste into disposable municipal waste using shredders that shred the waste into smaller particles which are then wetted with disinfectant spray and immersed in a disinfection solution. The wet waste is then dried using a hot off-gas in a drying chamber. Considering the number of clinics and hospitals in, middle-east, this method will be very successful if utilized, as it can treat the medical waste on-site which helps in better management of wastes.

- B. Current Medical Waste Treatment and Disposal in Developing Nations:** In developing nations, there are two key hospitals (public and private). Public hospitals represent 47.6%, while private and philanthropic hospitals represent 28.6% and 23.8%, respectively (Da Silva *et al.*, 2005). Manye *et al.*, (2003) revealed that in developing nations, incineration, open burning and burial of immunization waste (both infectious and non-infectious wastes) were practiced. A common practice is the disposal of infectious and regulated waste either in shallow pits (cemented or earthed) within the immunization office premises or in a locally developed incinerator. These observed immunization waste management practices are a risk to public health, especially when densely populated communities surround immunization offices. Studies revealed that collaborative efforts do exist between the immunization and health officers and the local government waste disposal board. Immunization and medical wastes are collected from the offices by the board and disposed of at government approved dumpsites or landfill sites. The only setback from this collaboration is that waste collected by this board is mixed with non-infectious waste and dumped at the site. The implication of this practice is that all of the wastes are now contaminated and could be classified as infectious waste. Other notable major problems in all of the surveyed sites were open dumping, burial or burning either within the immunization offices or at government recognised dumpsites, and non-treatment of infectious waste before final disposal.

Literature revealed that elsewhere, burning and burial of medical waste are practiced. Sadly enough medical infectious waste is not excluded from this practice. A common practice is the disposal of infectious and regulated waste types either on fallow land within hospital premises or in canal. This observed poor medical waste management practices is a risk to public health especially when most hospitals are surrounded by densely populated communities. In some countries in developing nations public hospitals are equipped with incinerators, which only being functional at the early years. In public hospitals with incinerators, infectious medical waste and sharps are incinerated in-house by the hospital's engineering department and engage the service of an environmental officer to oversee the treatment and eventual disposal of its medical waste.

In some developing nations, a collaborative effort does exist between public hospitals with facilities and some private hospitals for the incineration of their infectious and sharp waste types. This collaboration is guaranteed the quality of the environment. Some hospitals employ the services of private waste collectors for medical waste collection and final disposal at government authorized dumps. Notable among these private collectors is the state owned waste management authority that specializes in the collection of industrial wastes from most industrial set ups within the mega city. The major observed challenges to the environment are open dumping either within the hospital premises or at government recognized dumpsites; and non treatment of infectious waste before final disposal. Open dumps are known to have no controls over access of unauthorized persons or environmental pollution, hence, the potential health risk (WHO, 1999). Although, national policy on environment stipulates specific roles for appropriate government agencies to be among others; the determination and the use of environmentally safe and technologically sound techniques for disposal of toxic, hazardous and radio-active wastes in developing nations. Literature describes that various dumpsites within the vicinity of the metropolis do not meet necessary design criteria to operate as landfills for hazardous waste.

Medical waste even though, attracted a high level of segregation in most hospitals in developing nations. It is generally co-disposed with the municipal solid waste stream by waste handlers. The disposal of medical waste in municipal landfills therefore poses serious health and environmental hazards erstwhile landfills are mostly open dumps without adequate design consideration to guaranty protection of the environment from the disposal of such hazardous wastes. Treatment of infectious waste and sharps waste needs adequate legal backing and sanction in case of non-compliance by erring hospitals, health institutions and clinics. Leachates from beneath the landfill base may contain heavy metals and other organic pollutants that could lead to gross contamination of surface and groundwater resources (Longe and Williams, 2006).

Hazardous medical waste requires proper monitoring for an effective tracking at all times. Control of medical waste can only be fully achieved when adequate monitoring facilities are available. Control means that competent authorities can act rapidly to ensure the possibilities of minimizing inappropriate handling and dumping of medical waste. It also translates to mean that regulatory authorities should have the powers both legally and financially, to act quickly in order to reduce dangers posed to human health and the environment. In order to enhance adequate monitoring and control of medical waste in developing nations, governments at national and state levels need a comprehensive guideline on hazardous solid waste management other than what is currently in existence. For effective management therefore, there is need for cooperation among medical waste generators and environmental agencies with coordination among the various persons and corporate bodies involved in the control of medical waste activities. Also, handlers of medical waste within health care facilities should be trained on methods and new techniques of medical waste management and in hazardous effects of medical waste using WHO manual (WHO, 2005b; MINTER, 1979, CONAMA, 1993; COSEMA, 2000)

FUTURE RESEARCH DIRECTIONS

Research directions should include factors that can significantly improvements the current practices of handling and management of health care wastes in developing countries are required in order to ensure public health and environmental protection as well diverting significant quantities of material and resources from landfill.

CONCLUSION

It can be concluded that the:

- Medical waste generation varies from 0.17 to 8.4 kg/bed/day,
- Bulk density of the entire waste stream ranged from 137 to 359 kg/m³, with an average of 218 kg/m³,
- average bulk density for common (general) waste was 211 kg/m³, that for contaminated waste was 226 kg/m³ and that for special waste was 139 kg/m³,
- infectious (clinical, yellow bag) waste generation varied from 0.38 to 3.50 kg/bed-day;
- hospital waste generation (both infectious and mixed solid waste fractions) varies from 0.35 to 3.50 kg/ bed-day, and
- medical waste management practices varies from poor in developing nations to fair in developed nations

ACKNOWLEDGMENT

The authors wish to acknowledge the following people FEYIJIMI, O.O (CVE/2001/039);HASSAN, O.A (CVE/2001/040); OGinni, E.O (CVE/2001/053); OKAFOR, I.A. (CVE/2001/062); OLADELE, A.O. (CVE/2001/067), Salami G.T. (CVE/2001/082), (CVE/2002/078); CVE/2002/053; CVE/2002/031; CVE/2002/029; CVE/2002/015; CVE/2002/014 and CVE/2001/033 all of Department of Civil Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria for taking part in the study.

REFERENCES

- Abu Mhady, A. I., Awad, M. A., Al-Aghah, M. R., & El-Nahhal, Y. Z. (2019). Assessment of Medical Waste Management in El Shifa and Al Aqsa Hospitals- Gaza Strip. *Health, 11*(08), 1028–1042. doi:10.4236/health.2019.118081
- Askarian, M., Vakili, M., & Kabir, G. (2004). Hospital waste management status in university hospitals of the Fars province, Iran. *International Journal of Environmental Health Research, 14*(4), 295–305. doi:10.1080/09603120410001725630 PMID:15369994
- Bakiu, R., & Durmishaj, S. (2018). Medical Waste Effects and Management: Overview and Future Directions. *SF J Environ Earth Sci., 1*(2), 1019.
- CONAMA. (2001). *Resolution No. 283, National Environmental Council of Brazil*. Brazil Government Printing Office.
- COSEMA. (2000). *Resolution No. 09, State Environmental Council of Rio Grande do Sul*. RS Government Printing Office.
- Da Silva, C. E., Hoppe, A. E., Ravanello, M. M., & Mello, N. (2005). Medical wastes management in the south of Brazil. *Waste Management (New York, N.Y.), 25*(6), 600–605. doi:10.1016/j.wasman.2004.03.002 PMID:15993344

Diaz, L. F., Eggerth, L. L., Enkhtsetseg, S. H., & Savage, G. M. (2008). Characteristics of healthcare wastes. *Waste Management (New York, N.Y.)*, 28(7), 1219–1226. doi:10.1016/j.wasman.2007.04.010 PMID:17651963

EPA. (2002). *Environment Protection Act. Ministry of Environment*. Available from: <http://ncb.intnet.mu/eurd/minenv/epa.htm>

Glenn, M.R., & Garwal, R., (1999). *Clinical waste in Developing Countries. An analysis with a Case Study of India, and a Critique of the Basle- TWG Guidelines*. Academic Press.

Griffin, R. D. (1989). *Principles of Hazardous Materials Management*. Lewis Publishers Inc.

Hasan, M. M., & Rahman, M. H. (2018). Assessment of Healthcare Waste Management Paradigms and Its Suitable Treatment Alternative: A Case Study. *Hindawi Journal of Environmental and Public Health.*, 2018, 6879751. Advance online publication. doi:10.1155/2018/6879751 PMID:30151013

Lee, B.-K., Ellenbecker, M. J., & Moure-Ersaso, R. (2004). Medical waste management/incineration. *Waste Management (New York, N.Y.)*, 24, 143–151. doi:10.1016/j.wasman.2003.10.008 PMID:14761753

Lee, C. C., & Huffman, G. L. (1996). Medical waste management/incineration. *Journal of Hazardous Materials*, 48(1-3), 1–30. doi:10.1016/0304-3894(95)00153-0

Liberti, L., Tursi, A., Costantino, N., Ferrara, L., & Guzzo, G. (1994). Optimization of infectious hospital waste management in Italy: Part I – waste production and characterization study. *Waste Management & Research*, 12(5), 373–385. doi:10.1177/0734242X9401200502

Liberti, L., Tursi, A., Costantino, N., Ferrara, L., & Guzzo, G. (1996). Optimization of infectious hospital waste management in Italy: Part II. Waste characterization by origin. *Waste Management & Research*, 14(5), 417–431. doi:10.1177/0734242X9601400502

Longe, E. O., & Williams, A. (2006). A preliminary study of medical waste management in Lagos Metropolis, Nigeria, Iran. *Journal of Environmental Health Science & Engineering*, 3(2), 133–139.

Manyele, S. V., Anicetus, H., & Bilia, M. H. (2003). Globalization and its Effects on Medical Waste Management in Tanzania. *IET Annual Conference and General Meeting*, 76-92.

Marinkovic, N., Vitale, K., Afric, I., & Janev, H. N. (2005). Hazardous medical waste management as a public health issue. *Arhiv za Higijenu Rada i Toksikologiju*, 56(1), 21–32. PMID:15969205

Mato, R. R. A. M., & Kassenga, G. R. (1997). A study on problems of medical solid wastes in Dar es Salaam and their remedial measures. *Resources, Conservation and Recycling*, 21(1), 1–16. doi:10.1016/S0921-3449(97)00010-4

Meghdad, P., Younes, S., Mandi, D., & Diyari, A. (2016). Qualitative and Quantitative Survey of Waste Management System Case Study: Qods Hospital, Paveh-Iran (2015). *Eng. Applied Sci.*, 11(5), 1065–1065.

Minas, M., Spyridoula, G., & Dimitrios, K. (2017). Healthcare Waste Generation Worldwide and Its Dependence on Socio-Economic and Environmental Factors. *Sustainability*, 9(2), 220. doi:10.3390/u9020220

MINTER. (1979). *Act No. 53, Interior Ministry of Brazil*. Brazil Government Printing Office.

Miyazaki, M., Takuya, I., & Une, H. (2007). The treatment of infectious waste arising from home health and medical care services: Present situation in Japan. *Waste Management (New York, N.Y.)*, 25, 616–621. doi:10.1016/j.wasman.2005.01.003 PMID:16527476

Miyazaki, M., & Une, H. (2005). Infectious waste management in Japan: A revised regulation and management process in medical institutions. *Waste Management (New York, N.Y.)*, 25(6), 616–621. doi:10.1016/j.wasman.2005.01.003 PMID:15993346

Nemathaga, F., Maringa, S., & Chimuka, L. (2008). Hospital solid waste management practices in Limpopo Province, South Africa: A case study of two hospitals. *Waste Management (New York, N.Y.)*, 28(7), 1236–1245. doi:10.1016/j.wasman.2007.03.033 PMID:18060760

Oke, I. A. (2008). Management of Immunization Solid Wastes In Kano State, Nigeria. *Journal of Waste Management*, 28(12), 2512–2521. doi:10.1016/j.wasman.2007.11.008 PMID:18191394

PATH. (2005). *Training Health works in Management of Sharps waste, Guide for training of waste handlers*. www.ChildrensVaccine.org

Paweł, G. (2019). *Healthcare waste in the EU - a short overview*. Academic Press.

Rasheed, S., Iqbal, S., & Baig, L. A. (2005). Hospital waste management in the teaching hospitals of Karachi. *JPM. The Journal of the Pakistan Medical Association*, 55(5), 192–195. PMID:15960283

Remy, L. (2001). *Managing Hospital Waste is a Big, Nasty Deal*, Great Western Pacific Costal Post.

Soliman, S. M., & Ahmed, A. I. (2007). Overview of biomedical waste management in selected Governorates in Egypt: A pilot study. *Waste Management (New York, N.Y.)*, 27(12), 1920–1923. doi:10.1016/j.wasman.2006.08.009 PMID:17071069

Tsakona, M., Anagnostopoulou, E., & Gidarakos, E. (2007). Hospital waste management and toxicity evaluation: A case study. *Waste Management (New York, N.Y.)*, 27(7), 912–920. doi:10.1016/j.wasman.2006.04.019 PMID:16872822

Visvanathan, C. (2006). *Environmental Engineering & Management Program Asian Institute of Technology Thailand*. Asia 3R Conference, Tokyo, Japan. Access on <http://www.faculty.ait.ac.th/visu/>

WHO World Health Organization. (2004). Regional Office for Western Pacific and for South-East Pacific. *Practical Guidelines for Infection Control in Health Care Facilities*. SEARO and WPRO Publication No. 41.

Zarook, M. S. (2012). Medical Waste Management and Control. *Journal of Environmental Protection*, 3(12), 1625–1628. doi:10.4236/jep.2012.312179

ADDITIONAL READING

Ahmad, R., Liu, G., Santagata, R., Casazza, M., Xue, J., Khan, K., Nawab, J., Ulgiati, S., & Lega, M. (2019). LCA of Hospital Solid Waste Treatment Alternatives in a Developing Country: The Case of District Swat, Pakistan. *Sustainability*, 11(13), 3501. doi:10.3390/s11133501

Ambali, A. R., Bakar, A. N., & Merican, F. M. (2013). Environmental Policy In Malaysia: Biomedical Waste, Strategies And Issues. *The Journal of American Science*, 10(1), 1–17.

Bendjoudi, Z., Taleb, F., Abdelmalek, F., & Addou, A. (2009). Healthcare waste management in Algeria and Mostaganem department. *Waste Management (New York, N.Y.)*, 29(4), 1383–1387. doi:10.1016/j.wasman.2008.10.008 PMID:19091540

Cheng, Y. W., Sung, F. C., Yang, Y., Lo, Y. H., Chung, Y. T., & Li, K.-C. (2009). Medical waste production at hospitals and associated factors. *Waste Management (New York, N.Y.)*, 29(1), 440–444. doi:10.1016/j.wasman.2008.01.014 PMID:18359619

Coker, A., Sangodoyin, A., Sridhar, M., Booth, C., Olomolaiye, P., & Hammond, F. (2009). Medical waste management in Ibadan, Nigeria: Obstacles and prospects. *Waste Management (New York, N.Y.)*, 29(2), 804–811. doi:10.1016/j.wasman.2008.06.040 PMID:18835151

Dawar, A. B. K. (2017). *Hospital Waste Management Rules 2005 and Current Practices in Selected Hospitals of Peshawar Khyber Pakhtunkhwa*. Pakistan Abdul Basit Khan Dawar.

Glasser, H., Chang, D., & Hickman, D. (2012). An analysis of biomedical waste incineration. *Journal of the Air & Waste Management Association*, 1180–1188. PMID:1756039

Hossain, M. S., Santhanam, A., Norulaini, N. N., & Omar, A. M. (2011). Clinical solid waste management practices and its impact on human health and environment—A review. *Waste Management*, 31(4), 754–766.

Liu, H. C., You, J. X., Lu, C., & Chen, Y. Z. (2015). Evaluating health-care waste treatment technologies using a hybrid multi-criteria decision making model. *Renewable & Sustainable Energy Reviews*, 41, 932–942. doi:10.1016/j.rser.2014.08.061

Macaulay, B. M., & Odiase, F. M. (2016). Medical waste management practices in developing countries: A case study of health facilities in Akure, Nigeria. *International Journal of Environment and Waste Management*, 17(2), 103–127. doi:10.1504/IJEW.2016.076756

Malaysian Pharmaceutical Society. (2015, August 3). *National Pharmacy News*. Retrieved from <https://www.mps.org.my/index.cfm?andmenuid=107>

Marchetti, J. M. (2012). A summary of the available technologies for biodiesel production based on a comparison of different feedstock's properties. *Process Safety and Environmental Protection*, 90(3), 157–163. doi:10.1016/j.psep.2011.06.010

Mohammad, M. H., & Habibur, M. R. (2018) Assessment of Healthcare Waste Management Paradigms and Its Suitable Treatment Alternative: A Case Study. *Journal of Environmental and Public Health*. doi:10.1155/2018/6879751

Shamim Al Razib, M., Hasan, N., & Supriya, S. (2017). Study of Clinical Waste Management at Rajshahi Medical College Hospital (RMCH) in Bangladesh. *International Journal of Environmental Protection and Policy*, 5(2), 26. doi:10.11648/j.ijep.20170502.12

Tabasi, R., & Marthandan, G. (2013). A review on important factors in clinical waste generation rate. *Internatioanal Journal of Science and Technology*, 3(3), 194–200.

Teshiwal, D., Yazie, M., Girma, T., & Kasaw, A. C. (2019). Healthcare waste management current status and potential challenges in Ethiopia: A systematic review. *BMC Research Notes*, 12(1), 285. doi:10.1186/13104-019-4316-y PMID:31122274

Veronica, E. M., Osric, T. F., Linus, A., & Moforc, R. W. (2011). Health care waste management in Cameroon: A case study from the Southwestern Region. *Resources, Conservation and Recycling*, 57, 108–116. doi:10.1016/j.resconrec.2011.10.002

WHO. (2018). *WHO | Health-care waste*. Retrieved from <https://www.who.int/mediacentre/factsheets/fs253/en/>

Windfeld, E. S., & Brooks, M. S. L. (2015). Medical waste management – A review. *Journal of Environmental Management*, 163, 98–108. doi:10.1016/j.jenvman.2015.08.013 PMID:26301686

Windfeld, E. S., & Brooks, M. S.-L. (2015). Medical waste management – A review. *Journal of Environmental Management*, 163, 98–108. Advance online publication. doi:10.1016/j.jenvman.2015.08.013 PMID:26301686

Yuwono, A. S. & Ersu, N.S.(2018) Evaluation of Medical Solid Waste Management: A Case Study of Two Hospitals in Bogor. *Indonesia International Journal of Applied Environmental Sciences*, 13(3), 323-337.

KEY TERMS AND DEFINITIONS

BHW: Biohazardous wastes.

EPI: Expanded Program on Immunization. The Expanded Program on Immunization is a World Health Organization program with the goal to make vaccines available to all children

IV: An intravenous central line is a kind of intravenous (IV) line used to give medicines and fluids. It is a thin, soft, plastic tube called a catheter that is inserted through the skin and into a vein.

MW: Medical waste.

MWC: Medical waste commission.

MWM: Medical waste management.

MWMP: Medical waste management plan.

RCW: Recyclable wastes.

SW: Sharp wastes.

TP: Training program.

WHO: World Health Organization.

Managing Water Resources: Industry Initiative

10

Sabyasachi Nayak

Confederation of Indian Industry, India

INTRODUCTION

Water is inextricably linked to every aspect of life and development. The burgeoning population leading to spike in anthropogenic activity and climate change has put enormous strain on the fragile water resources. Continuing with the current water trajectory, country like India has witnessed a progressive decline in per capita availability. Numerous steps are initiated by the government to address the issue plaguing the water sector in India. But the involvement of various stakeholders is equally important.

The Indian industry has taken a proactive role in ensuring the proper management of the limited water resources by engaging with concerned stakeholders. This paper seeks to capture the grass roots interventions undertaken by industry in improving the management of water resources. As water can directly impact the livelihood therefore the conservation, revival and development of water resources calls for a comprehensive, holistic, scientific and sustainable approach. It puts into perspective the operational modality of industry in engaging with the community to manage water resources. The objective is to provide a dynamic insight into the operational model of the industry in addressing the accumulated neglect in the water sector. An assessment of the operating model reiterates the positive social, economic and environmental outcome in a sustainable manner. Therefore, it is proposed to explore operating beyond the level of “watershed” to larger level “river basin”. Going forward it is imperative to revive a culture of community base management of natural resources.

BACKGROUND

Water use has been increasing worldwide by about 1% per year since the 1980s. (United Nations Educational Scientific and Cultural Organization [UNESCO], 2019) Global water use has increased by a factor of six over the past hundred years. Levels of physical water stress are likely to increase as populations and their demands for water grow, and the effects of climate change intensify (UNESCO, 2018). The planet’s capacity to sustain the growing demands for freshwater is being challenged and by 2050, global water demand is projected to increase by 55% (UNESCO, 2017). Declining water resources are already affecting many parts of the world and an estimated 20% of the world’s aquifers being over-exploited (UNESCO, 2014). The gap between water supply and demand is projected to be of 40% in 2030. Asia will need on average 65% more freshwater withdrawals for their industry and energy sectors by 2030 in order for their national economies to grow as forecast (Water Resource Group, 2012). India is among the ten major water users in the world in terms of volume, which uses 646 km³ of water a year (UNESCO, 2009).

DOI: 10.4018/978-1-7998-3479-3.ch092

Water Resources: India

India's finite and fragile water resources are stressed and depleting, while sectorial demands (including drinking water, industry, agriculture, and others) are growing rapidly in line with urbanization, population increases, rising incomes and industrial growth (Government of India, 2001). The average rainfall is about 1170 mm which varies temporally and spatially (Government of India, 2015). About 80% of the country's surface water is polluted. The rivers Yamuna, Ganga, Gomati, Ghaghara, Chambal, Mahi and Vardha are amongst the most coliform-polluted water bodies in India. (WaterAid, 2015). India is the largest user of groundwater in the world with an annual extraction of 210 billion cubic meters. As per NASA study groundwater depletion in Rajasthan, Punjab, Haryana and Delhi was equivalent to a net loss of 109 cubic km from August 2002 to October 2008 (Government of India, 2010).

The irrigation sector consumes about 80% of the total water use which may reduce to about 70% by 2050, whereas the domestic sector and industrial consumption is projected to double during the same period (Government of India, 2014). India faces an unprecedented crisis in the next two decades. Today, major urban areas are unable to provide a reliable and regular water supply. Not surprisingly, aquifer depletion and inefficient water use are now endemic (UNESCO, 2012). The demands of a rapidly industrializing economy and urbanizing society come at a time when the potential for augmenting supply is limited, water tables are falling and water quality issues have increasingly come to the fore. It is no wonder then that conflicts across competing uses and users of water are growing by the day (Government of India, 2011).

Industry Water Initiative

In the advent of water crisis, the industries operating in the country will be the hard hit. Therefore the Indian industries collectively have taken a proactive role in ensuring the proper management of the limited water resources (Confederation of Indian Industry, 2006). The embracing challenge for industry is to play its appropriate role in effectively addressing unsustainable exploitation and contamination of freshwater resources (UNESCO, 2012). Industry has been showing ways of effectively managing the scarce water resource through proper engagement, networking and partnering with stakeholders (Confederation of Indian Industry, 2013). Industry is continuously seeking better ways to manage its water resources efficiently (Associated Cement Companies, 2013). Using a participatory development approach, construction and revival of water harvesting structures including Revival of village Ponds and Pasture land development is ensuring available water pre and post rainy season for drinking, cattle and irrigation purposes. Community is empowered to take local ownership of Water Projects and actively engages in creating solutions to water challenges .(Ambuja Cement Foundation, 2017).

Watershed Development Programme, designed to assist farmers in identified moisture-stressed has improved the water situation. The focus is not only seeking to achieve water conservation and soil enrichment, but also improving village-based management of water and other natural resources by evolving a culture of optimum water usage, finally leading to maximizing the benefits of water resources. (ITC Limited, 2017) Measures such as construction of check dams, ponds, excavation of streams, farm bunds for soil and water conservation practices enhances the availability and accessibility of natural resources on which rural livelihoods and economy depend (Lupin Human Welfare and Research Foundation, 2016). Rejuvenation of Rivers and Streams have proved extremely useful to convert rain fed farming into irrigated thereby improving the crop productivity. The intervention of Stream Revival mainly involves the excavation of the river beds and creating embankments to help channelize the water from the natural

Stream/ Rivers. The Check Dams are constructed across the river banks and Streams. This helped in increasing water storage in stream which was utilized latter for the irrigation purpose and it also helps augment ground water level in surrounding area (Kamalnayan Jamnalal Bajaj Foundation, 2017).

Water harvesting structures constructed serves as a critical source of irrigation for communities living in dry and rain-dependent areas. By implementing soil and water conservation practices communities have managed to substantially mitigate the impact of excessive rainfall as well as drought. [<https://www.reliancefoundation.org/water-security>] This has brought not only more cultivable area under irrigation but increase in cropping intensity from a single crop to double crops (Lupin Human Welfare and Research Foundation, 2017) . Similarly, sub-surface dykes and run-off diversion systems are able to provide water for irrigation for longer duration in a year, creating a positive impact on agriculture yields, as well as increasing accessibility of water for domestic use in water scarce areas. Salinity ingress in coastal areas is addressed through promotion of low-water intensive crops, renovation of ponds, and deepening of wells to recharge ground water resources (Ambuja Cement Foundation, 2015). In order to ensure effective use of water resources and raise awareness around water security issues, water budgeting is emphasized upon (Reliance Foundation, 2016). Training on modern agricultural practices to improve productivity includes drip irrigation, intercropping and vermicomposting (Asian Paints Limited, 2015).

FOCUS OF THE ARTICLE

The article focuses on the operational model of industry at grass roots level for managing water resources and a contemporaneous analytical assessment of the same. It examines how the industry beyond its “industrial operation” has partnered with community, developed requisite capacity of community and through robust monitoring & evaluation has brought about a desired change in the water scenario.

Operating Model

The model of operation resorted to by various industry for water management can be clubbed into two categories

- Direct implementation by industry through its Foundation/ Trust: Industry engages directly with the community through their trained and professional staff with the community.
- Implementation through NGO partner: Industry merely funds the initiative and the implementation responsibility is totally entrusted with the NGO

In both these models partnership with the community remains central and the intervention is projectivized – baseline Survey, assessment of Needs, proposed action, earmarking funds under each activity, time line, deliverables, mode of implementation, monitoring, documentation and exit strategy.

In the contemporary context forging partnership between Government, industry and community is critical to stimulate, strategize and synchronize the planning, development and management of water resources through a common integrated perspective. Such effort in a systematic and sustained manner has led to the evolution of multistakeholder partnership, incorporating a Public Private Community Partnership (PPCP) approach. The essence of PPCP in water related projects has been to pivot engagement between private sector and community in a synergetic operational mode to embark towards water security. The role of various stakeholders in multistakeholder PPCP is captured in table 1.

Currently, industry is implementing water related projects for the community adhering to the concept of watershed management which spans for five to six years. The overall management of watershed initiative is interdisciplinary with representation from hydro geologists, civil engineering, environmental science, hydrology, agronomist, veterinary and more importantly the social scientists. Harnessing the skills, expertise and competency of respective discipline in an integrated manner plays a pivotal role in influencing the outcome and impact of the initiative. The implementation of the project envisages a responsive governance mechanism, effective coordination between multiple partners, timely availability of funds and execution at the grass roots level to improve holistic water management. At the community level, Village committee embracing representation from community based organizations, youth, self-help groups, marginalized group galvanize the process of implementation in a participatory manner.

A robust monitoring & evaluation mechanism is put in place to provide improvement oriented direction, guidance and strategies in a participatory manner for introducing course correction measures and facilitate enhanced learning from ground realities. Participatory monitoring & evaluation conducted at the grass roots level assess progress, evaluate performance, and identify implementation issues. This has led to the evolution of a developmental process galvanizing action and results owned by the beneficiaries of the project. Monitoring & Evaluation is undertaken at various scale- day to day monitoring,

Table 1. Role of Stakeholders in PPCP

Role	Public	Private		Community	
	Government	Industry	NGO	Village Committee	Farmer
Capital Funding of the Project		√			
Technical expertise	√	√			
Community Mobilization		√	√	√	
Sensitization of Stakeholders					
Preparation of technical details					
Site Selection for construction	√		√	√	
Community interface			√	√	
Define Project Deliverables			√	√	√
Grass roots implementation			√	√	
Adherence to technical details	√	√	√		
Monitoring	√		√	√	
Capacity building on water use efficiency			√		
Operation & Maintenance of structures created				√	√
Demonstration of Water smart Agricultural Practices		√	√		√
Improved Water use efficiency			√		√
Crop diversification			√		√
Artificial recharge of wells			√	√	
Monitoring of wells			√	√	√
Water treatment			√	√	
Equitable distribution of water				√	√
Dispute Resolution				√	

periodic monitoring, concurrent progress monitoring, process monitoring and result monitoring. The overall Monitoring & Evaluation framework tracks the inputs, outputs, outcomes and impacts, thereby strengthening the data systems for better collection, collation, and analysis for tracking progress. A user friendly MIS system facilitates receiving systematic inputs from partner's to track progress- physical, financial, output and outcomes. Thus it helps stakeholders in availability of real time credible and reliable data to arrive at strategic decision to refine further plan of action and uphold transparency in operation by reinforcing accountability.

Requisite capacity development exercises taking into account the existing capacity and addressing areas to achieve desired capacity with the involvement of stakeholders and particular focus on community has led to improvement in functioning of activities at grass roots level. This quantitative and qualitative enhancement of human resources in applying their acquired skills, competency and knowledge has been instrumental in affecting the efficacy of the institutions at the grass roots level in undertaking planning, implementation, supervision and monitoring of initiative to improve water scenario. Technical competency focuses on implementation of integrated water resources management, improved agricultural practices, water budgeting, enhancing water use efficiency and water quality protection.

A responsive feedback mechanism at the grass roots in addressing the issues of the community has been vital in ensuring their active participation. The newly created extension service and field staff act as an interface between community and industry. In few project area users friendly WhatsApp has made a marked difference in terms of availing timely information, posting of pictures or sharing voice recording thereby establishing connectivity amongst stakeholders.

Assessment of the Operating Model

Collaborative, result oriented and focused partnership is the distinctive feature that has demonstrated in reaching out to varied stakeholders in harnessing their comparative advantages. Over the years, these models have been refined to reach the current form, which can be emulated by other industry while venturing into similar venture. The learning's emanating is essential in reinforcing efficacy of the intervention and evolving benchmarks for the implementation of projects at the ground level. In spite of the efficacy of the model a deep sense of skepticism pervades among other stakeholders, particularly the farming community whenever industry embarks on managing water resources. Therefore, gaining the confidence of the community and other stakeholder is a daunting task for the industry.

Industry has demonstrated its proactive role and earnestness to represent the need to seriously address the burning issue of water. Improving the water management has also ensured "ecosystem security". In overall terms, the positive social, economic and environmental outcome of the initiative has ensured a win situation for the varied categories of stakeholders. Transformation in the water sector is attributed to the creation of trained manpower, mobilization of resource and bringing about a mental revolution in management of the depleting scarce resource. In spite of this, funding has remained the fulcrum of partnership with industry thereby downplaying their managerial abilities, knowledge and expertise.

The essence of community ownership in terms of instituting, implementing and being accountable for the initiative is gaining momentum. This ecosystem has ripple effects on various aspects impacting the quality of lives of the community. With water as the central plank to development, the other areas where positive impact witnessed are reflected in table 2. The model re-cast the debate on water management in the contemporary context leading to development. There has been a positive shift from "dependency" to "development" which is realized by ensuring continuous, sustained and systematic engagement with the community.

Table 2. Outcome of Water Intervention

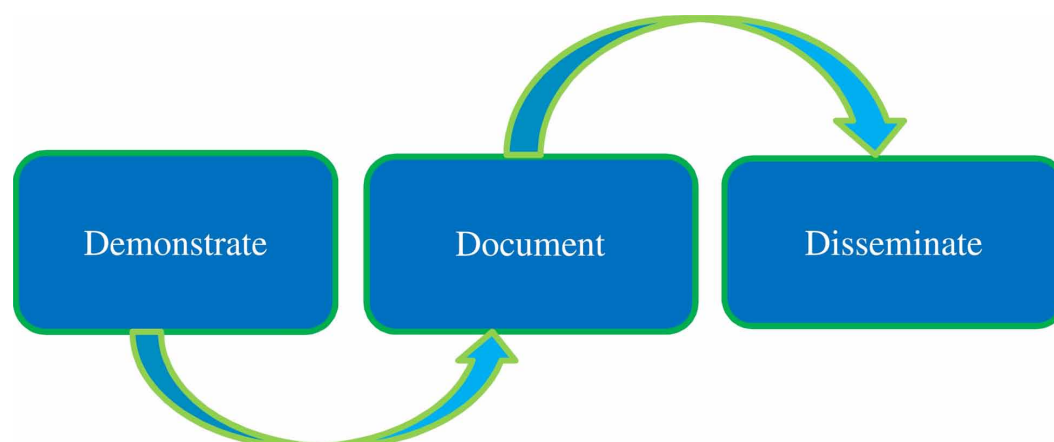
Social	Environmental	Economic
-Expand enrollment of girl child -Women participation in decision making -Curtailement in Household Migration -Reduced drudgery of women -Debt free farmers -Functional Community Organization	-Strengthen management of natural resources -Protection of ecosystem -Improvement in flora and fauna -Recharge of Ground water -Creation of water storage -Conjunctive use of surface and ground water -Improvement in water quality	-Increase in production and income -Diverse economic activity -Employment Opportunities created -Reduction in farm input cost -Slash in treating water borne disease

SOLUTIONS AND RECOMMENDATIONS

A serious multistakeholder partnership operating beyond the level of “watershed” to larger level “river basin” can make an impact in addressing the pressing issue of water management. **Figure 1** emphasizes 3D approach-“*Demonstrate Document Disseminate*” to venturing into “river basin” for implementation of water initiatives. The initiative should be cognizant to demonstrate that the planning, development and management of the scarce, fragile and finite water resources is undertaken in a unified manner without neglecting the “unity & integrity of the hydrological cycle”. It is equally important to documents the entire process for gaining an insight into the nuance and dynamics relating to various aspects of the project implementation. Furthermore, the documentation should be widely disseminate so that other stakeholder venturing into similar endeavor adopt the recommended approach and in the process need not reinvent the wheel.

As the water crisis looms across large parts of the country, therefore an amalgamation of scientifically rigorous orientation, participatory community approach, application of cost effective technology, demand management strategies, reinforcing improve water usage efficiency and resilience building can bring about a paradigm shift in water management. Nonetheless, bringing about behavioral changes will be instrumental in ensuring that the various activities undertaken are in consonance with the hydrological cycle. There is an urgent need for institutionalizing partnership between Government and industry, whereby desired impact can also be achieved in Government water management programs. Industry can also play a pivotal role identifying gap that deters the accrument of desired impact of government program.

Figure 1. 3D approach-demonstrate document disseminate



FUTURE RESEARCH DIRECTIONS

Community has the requisite wisdom but it needs to be harnessed in a scientific manner for bringing about the desired change. Active community participation, ownership and accountability are central to the success of water intervention. Therefore, future research may include a more thorough exploration of reviving effective community base natural resource management. It is imperative for the community to be the custodian, guardian and pride themselves on being steward of the natural resources. A mechanism making grass roots institution functional with requisite capacity and transparent operation can help in reinventing their water trajectory for the future.

CONCLUSION

Water management can be ensured through active participation and involvement of all stakeholders. In this direction industry has taken a proactive role by emphasizing the centrality of community participation. The prevalent operating model of the industry can be categorized into two parts -direct implementation by industry through its Foundation/ Trust; and implementation through NGO partner, focusing on- partnerships with community; robust monitoring & evaluation mechanism; requisite capacity development; and a sensitive feedback mechanism. With water as the central plank for development the ripple effects have been witnessed in other areas- agriculture, livelihood, education, health, poverty and empowerment. Thus, the operating model has been able to displace the entrenched “dependency” syndrome with “development”.

It is time to reckon multistakeholder partnership operating beyond the level of “watershed” to larger level “river basin” for making an impact in addressing the pressing issue of water management. Adhering to the 3D principle-“*Demonstrate Document Disseminate*” industry can contribute in creating the desired impact in other project settings. Simultaneously, research endeavor aimed at resuscitation and reassertion of community as the custodian of natural resource will be crucial. Going forward it is imperative to take a multidisciplinary stratagem, adopt participatory approach, and continuously strive for innovation and ingenuity thereby prioritizing local specific and customized solutions.

REFERENCES

- Ambuja Cement Foundation. (2015). *Annual Report 2014-15*. Mumbai: Author.
- Ambuja Cement Foundation. (2017). *Annual Report 2016-17*. Mumbai: Author.
- Asian Paints Limited. (2015). *Movement & Improvement: Sustainability Report 2014-15*. Mumbai: Author.
- Associated Cement Companies. (2013). *Sustainable Development Report 2013*. Author.
- Confederation of Indian Industry. (2006). *PAANI*. Author.
- Confederation of Indian Industry. (2013). *Moving towards strategizing and mainstreaming CSR initiatives of Indian industry in water sector- building partnerships and bridging the gaps*. Author.
- Government of India. (2001). *India Assessment Report*. Planning Commission, Government of India. New Delhi: Author.

- Government of India. (2010). Midterm Review. Planning Commission. New Delhi: Author.
- Government of India. (2011). Faster, Sustainable and More Inclusive Growth: An Approach to the Twelfth Five Year. Planning Commission. New Delhi: Author.
- Government of India. (2014). *Guidelines For Improving Water Use Efficiency in Irrigation Domestic & Industrial Sector*. Central Water Commission. Ministry of Water Resource. Author.
- Government of India. (2015). INDC- Vulnerability Report. Ministry of Water Resources, River Development and Ganga Rejuvenation National Water Mission. New Delhi: Author.
- Kamalnayan Jamnalal Bajaj Foundation. (2017). *Annual Review 2016-17*. Author.
- Limited, I. T. C. (2017). Sustainability Report 2017. Kolkata: Author.
- Lupin Human Welfare and Research Foundation. (2016). *Annual Report 2016*. Author.
- Lupin Human Welfare and Research Foundation. (2017). *Enabling Paradigm Shift in Rural Living: Annual Report 2016-17*. Mumbai: Author.
- Reliance Foundation. (2016). *Annual Report 2015-16*. Mumbai: Author.
- United Nations Educational Scientific and Cultural Organization. (2009). *The United Nations World Water Development Report 3: Water in changing world*. Author.
- United Nations Educational Scientific and Cultural Organization. (2012). *The United Nations World Water Development Report 4: Knowledge Base, Volume-II*. Author.
- United Nations Educational Scientific and Cultural Organization. (2014). *The United Nations World Water Development Report 2014: Water and Energy*. Author.
- United Nations Educational Scientific and Cultural Organization. (2017). *The United Nations World Water Development Report 2017. Wastewater: The Untapped Resource*. Author.
- United Nations Educational Scientific and Cultural Organization. (2018). *The United Nations World Water Development Report 2018: Nature-Based Solutions for Water*. Author.
- United Nations Educational Scientific and Cultural Organization. (2019). *The United Nations World Water Development Report 2019: Leaving No One Behind*. Author.
- Water Resource Group. (2012). Background, Impact and the Way Forward. Water Resource Group-2030. Author.
- WaterAid. (2015). *Urban WASH: An Assessment of Faecal Sludge Management (FSM) Policies and Programmes at the national and select states level*. Author.

ADDITIONAL READING

- Confederation of Indian Industry. (2010). *PPCP Guidebook on Water*. Author.
- Confederation of Indian Industry. (2010). *Our Cup of Joy*. Author.

Confederation of Indian Industry. (2010). *Breaking the Boundaries in Water Management: A Case Study Booklet*. Author.

Confederation of Indian Industry. (2013). *Moving Towards Strategizing and Mainstreaming CSR Initiatives of Indian Industry in Water Sector- Building Partnerships and Bridging the Gaps*. Author.

Confederation of Indian Industry. (2018). *Maharashtra Annual Report 2017-18*, Mumbai: Author.

Nayak, S. (2011). *PPCP Initiative for Water Management*. National Institute of Hydrology.

Nayak, S. (2014). Synchronizing Indian CSR Endeavor of Industry. *International Journal of Civic Engagement and Social Change*, 1.

Nayak, S., & Sharma, S. (2012). *Industrial Association Initiative for Water Resource Management*. Transformacje.

KEY TERMS AND DEFINITIONS

Hydrological Cycle: The origin, movement, and distribution of water in the planet.

Innovation: The process of evolving a new idea or concept from an existing practice bringing about a transformation in addressing a pressing issue.

Operating Model: The systematized approach adhered to by industry in ensuring delivery of requisite development outcome leading to improvement in the quality of lives.

Public Private Community Partnership: This deals with engagement of government, private sector, and civil society to achieve the aims of policy and ensure policy implementation.

Stakeholder: Stakeholders include individuals, groups, and institutions with an interest or concern in an endeavor.

Sustainability: Utilize, maintain, and conserve ecological balance by avoiding depletion of natural resources (in this case water) so that the demands of the present generation are met, simultaneously ensure that the concern of the future generation is not compromised.

Water Management: The process of planning, developing, and managing water resources in a holistic manner without compromising the needs of the future generations.

Water Security: Ensuring the availability and accessibility of reliable, safe, and sufficient amount of water in a sustainable manner.

Inhibited Antibiotic-Resistant and Electrochemical Treatment of Pharmaceutical Wastewater

Isaiah Adesola Oke

 <https://orcid.org/0000-0002-7082-7682>

Obafemi Awolowo University, Nigeria

Fehintola Ezekiel Oluwaseun

Adeyemi Federal University of Education, Ondo, Nigeria

Justinah S. Amoko

Adeyemi Federal University of Education, Ondo, Nigeria

Salihu Lukman

 <https://orcid.org/0000-0002-6567-1257>

University of Hafr Al-Batin, Saudi Arabia

Adekunbi Enoch Adedayo

Adeyemi Federal University of Education, Ondo, Nigeria

INTRODUCTION

Water is vital both as a solvent in which many of the body's solutes dissolve and as an essential part of many metabolic processes within the human body. Potable water production and supply are essential ingredients in the community and to human beings and other living things. Waterborne diseases (diarrhoea, typhoid, and cholera) that are widely spread in developing countries and in sub-Saharan Africa are caused mainly by poor sanitation and hygiene practices, which involve collection, treatment and disposal of solid and liquid wastes. In wastewaters (liquid wastes) microorganisms are removed by disinfection treatment processes using Chlorine dioxide, Chlorine compounds (monochloroamine, dichloroamine), Ozone and ultra-violet radiations. The stringent potable water standard and effluents quality limits by Environmental Regulatory Authorities, increasing energy cost and reports of outbreak of water-borne diseases in many parts of Africa have brought a greater challenge to Water experts (Water Science, Environmental and Water Engineers). These actions have caused Environmental and Water Engineers to look for effective methods of treating water and wastewaters rather than relying on conventional water and wastewater treatment methods.

In the last two decades several studies have been conducted on the performance of electrochemical treatment process as a wastewater treatment method. The electrochemical treatment process has been effectively used in the treatment of various industrial, municipal, domestic- institutional and domestic wastewaters. Wastewater generated from industrial processes such as distillery industry (Manisankar *et al.*, 2004), textile industries (Naumczyk *et al.*, 1996) and dye processing industry had been treatment using electrochemical treatment process. Chen (2004) presented documents on the design development,

DOI: 10.4018/978-1-7998-3479-3.ch093

construction and applications of electrochemical techniques in water and wastewater treatment. In the documentation a particular focus was given to selected techniques such as electrodeposition, electrocoagulation, electroflotation and electrooxidation with little data on treatment of wastewaters from pharmaceutical industries. Over 300 related publications were reviewed with 221 cited or analysed. Literature (Li *et al.*, 2014, Zheng *et al.*, 2015) have provided more information on efficacies of various electrochemical treatment of water and wastewaters. The main focus of this study is to conduct a simple literature review on treatment wastewaters from pharmaceutical and related industries and establish efficacies of electrochemical treatment technique in removing selected pollutants form raw water.

BACKGROUND

Carbon resin electrodes were developed from used dry cells (Oke *et al.*, 2007a, b and c). Electrolysing equipment was developed from local materials. Synthetic (simulated) wastewaters were prepared using procedures and methods specified in APHA (1998). Fractional factorial (2^{K-P-1}) experiments were utilized at random to determine influence of selected factors (separation distance between the electrodes, volume of the wastewater used, applied current, temperature of the wastewater, treatment time, concentration of the pollutant, concentration calcium of hypochrite ($\text{Ca}(\text{OCl})_2$) added and depth of the electrode into the wastewater used) on efficiency of electrochemical process in removing Biochemical Oxygen Demand concentration at five days (BOD_5) as follow-up on previous study (Oke *et al.*, 2007c). The choice of the parameters to be studied was done on the basis of the theoretical data about several factors that determine the efficiency of an electrochemical method and the scarce knowledge concerning carbon-resin / aluminium electrodes. An electrochemical treatment plant on a laboratory scale using public electricity source was setup (Figures 1a and b). BOD_5 determinations in wastewaters were carried out following procedures in APHA (1998) using respirometric method (CAMLAB HACH, model number 2173B BOD manufactured by Hach Chemical Company). The procedures and steps were repeated for blanks (to serve as controls). Efficacies of the system were computed using equation (1) as follows:

$$Y = 100 \left(\frac{(C_0 - C_t)}{C_0} \right) \quad (1)$$

Where; Y is the efficacy of the system (electrochemical treatment process), C_0 is the initial BOD_5 and C_t is the final BOD_5 . Table 1 presents arrangement of Fractional factorial (2^{K-P-1}) experiments. Effects of the selected factors on the efficacy of the electrochemical treatment process were computed using contrast method (Guttman *et al.*, 1971; Gardiner and Gettinby, 1998; Davore and Farnum, 1999, Davore, 2000) as stated in equation (2) as follows:

$$ef = \left(\sum_{i=1}^{K-P-1} l_i Y_i \right) \quad (2)$$

Where; K is the number of the factors (8), P is the fraction of the factorial experiment (3), l_i is the level (-1 or 1), and ef is the effect of the factor. Number of experiments required for both factorial and fractional factorial were computed using the following equations:

$$N_{ffn} = 2^{K-P-1} \quad (3)$$

Where, N_{ffn} is the number of experiments required for fractional factorial experiments. Statistical evaluations were conducted using the following equations as follows:

$$SSQ = ef^2 \times (2^{K-P-2}) \quad (4)$$

Where; SSQ is the sum of square of the factor

$$MSS = \left(\frac{SSQ}{df} \right) \quad (5)$$

$$SST = \sum_{i=1}^N Y_i^2 - \left(\frac{\left(\sum_{i=1}^N Y_i \right)^2}{N} \right) \quad (6)$$

$$SSE = SST - \left(\sum_{i=1}^{Nn} SSQ_i \right) \quad (7)$$

$$F = \left(\frac{MSS}{MSE} \right) \quad (8)$$

Where; SST is the sum square of total, SSE is the sum square of error, Nn is the number factors evaluated, MSS is the mean sum of square, MSE is the sum of square of the error, df is the degree of freedom (number of levels – 1) and F is statistical F- value.

Industrial wastewaters were collected from selected industrial sources according to APHA (1998) within the country (Nigeria). The industrial wastewaters collected were subjected to electrochemical treatment and chemical treatments individually and in combinations under specified treatment conditions.

Effects of the selected treatment (electrochemical, combination of electrochemical and chemical, modified electrochemical) in removing selected contaminants (chloride ion, BOD₅, Pb, Cd, suspended solid and total solids) from typical pharmaceutical wastewaters were monitored. In the analysis of selected pollutants concentrations, procedures in (APHA (1998) were followed.

RESULTS AND DISCUSSION

Results from this study are presented and discussed in three categories (literature review, treatment of simulated pharmaceutical wastewaters and treatment of typical pharmaceutical wastewaters) as follows:

Literature Review

10

Emerging pollutants (EPs) are defined as synthetic or naturally occurring chemicals that are not commonly monitored in the environment but which have the potential to enter the environment and cause known or suspected adverse ecological and (or) human health effects. EPs are categorized into more than 20 classes related to their origin (Figure 2). The prominent classes are: pharmaceuticals, pesticides, disinfection by-products, wood preservation and industrial chemicals. There are more than 1036 emerging pollutants, their metabolites and transformation products, are listed as present in the European aquatic environment (Geissen *et al.*, 2015 ; NORMAN, 2019). In light of the potential impact of these substances on aquatic life and human health, the lack of knowledge regarding their behaviour in the environment and the deficiency in analytical and sampling techniques, action is urgently required at multiple levels. With reference to the definition, EPs are wastewaters from pharmaceutical industries. Figure 3 provides various concepts for management of emerging pollutants in the aquatic environment. Physicochemical mechanisms can also contribute to EP removal from wastewaters and surface waters. Melanie *et al.* (2019) reported that the degradation of Ciprofloxacin, enoxacin, levofloxacin, lomefloxacin and norfloxacin by oxidation processes such as aeration can be described by first order kinetics. Figure 4 provides multi-scale assessment of the ecological and human health risks derived from the acute and chronic exposure to EPs contained in surface water. Figure 5 presents photoinduced degradation of Ciprofloxacin, enoxacin, levofloxacin, lomefloxacin and norfloxacin. Best *et al.* (1960) show that efficacy of disinfectant should be scrutinised regularly to assure proper activity against major medically significant organisms. It was documented that most of the formulations studied and findings emphasized the need for caution in selecting an appropriate disinfectant for use on contaminated surface water bodies, particularly in the presence of organic material.

Martin *et al.* (1966) reported that iodine appeared to be more effective than chlorine against both the standard faecal indicators. It was documented that against *staphylococci* derived from the mouth, nose, and skin iodine appeared to be more effective than chlorine. Some microorganisms inherently developed resistance to disinfectant and even grow in disinfectant solutions. Frank *et al.* (1969) documented that resistant cells of *Pseudomonas aeruginosa* and waterborne *Pseudomonas sp.* were able to multiply in nitrogen-free minimal salts solution containing various concentrations of commercially prepared; ammonium acetate buffered benzalkonium chloride (CBC), which is a potent antimicrobial agent.

It was reported that severe problem of release of antibiotic or disinfectant resistant bacteria to the environment is potential of transferring of the resistant gene to community bacteria. In Philadelphia, examination on the transfer of plasmids pBR322 and pBR325 in wastewater from laboratory strains of *Escherichia coli* to bacteria indigenous to the waste disposal system revealed that bacterial strains isolated from raw wastewater or a plasmid free *E. coli* laboratory strain served as recipients. The transfer of the pBR plasmid into the recipient strain occurred within 25 hours co-incubation in either L broth or sterilized wastewater (Michael *et al.*, 1985). It was reported that after the co-incubation recipients exhibited both plasmid-encoded phenotypic characteristics and an altered plasmid profile, which are shown by agarose gel electrophoresis of purified plasmid DNA (Michael *et al.*, 1985). In France, comparison between *in vitro* activity of Cefoxitin with Metronidazole and Clindamycin against 322 strains of anaerobic bacteria collected from several hospitals. The antibiotic-resistant were tested using an agar dilution method. The result showed that Metronidazole and Cefoxitin inhibited at least 89% of antibiotic-resistant strains tested. Clindamycin was found to be less active in inhibited antibiotic-resistant (Dubreuil *et al.*, 1984).

Roberto *et al.* (2005) reported that amoxicillin is a widely used antibiotic, which has been found in effluents from Sewage Treatment Plant. Evaluation of ozonation treatment of pharmaceuticals was

proposed as a process for its abatement from these pharmaceuticals effluents. The results of ozonation treatments of amoxicillin containing wastewaters indicated that ozone attack was mainly directed towards the phenolic ring of the studied molecule leading to the formation of hydroxyl-derivative intermediates. In addition the study revealed that there was no strong evidence of attack of ozone on sulphur atom with sulphur-oxide formation. Pauwels and Verstraete (2006) reviewed different techniques in use for hospital generated wastewater treatment and evaluated actions of technologists and ecotoxicologists as well as public health specialists.

In Aboulhassan *et al.* (2006) reported that coagulation and flocculation with FeCl_3 proved to be effective in a pH range between 7.00 and 9.00. The process reduced surfactants and COD by 99.00% and 88.00% respectively, and increased BOD_5 to COD ratio from 0.17 to 0.41. It was concluded that precipitation coagulation process, adsorptive micellar flocculation mechanism contributed to the removal of surfactants and organic matters from the wastewaters.

Zhimin *et al.* (2006) revealed that aeration significantly improves lagoon functionality, by increasing both organic and ammonia removal. Ammonia present in the wastewaters played a critical role in antibiotics disintegration and bacterial reduction due to conversion of free chlorine to monochloramine and dichloroamine. Commonly, a chlorine dose close to breakpoint chlorination is required to achieve effective removal of antibiotics. This action usually leads to higher consumption of free chlorine than necessary in most of the wastewaters treatment.

Most of the antibiotics taken by the patients are partially metabolized and excreted through faeces and urine. Residual quantities of these antibiotics in faeces and urine reached the wastewater. The residual quantity of antibiotics exposes the bacteria in hospital wastewaters to a wide range of biocides that could act as a selective pressure for the development of resistance (Nuñez and Moretton, 2007). Increasing attention has been directed recently to the resistance of bacteria to antibiotics and disinfectants. The resistant bacteria isolated were diverse in nature. In Buenos Aires City hospital, Brazil, the bacterial population resistant to disinfectants were *Enterobacteriaceae*, *Staphylococcus spp.*, and *Bacillus spp.*, which are highly associated to nosocomial infections (Nuñez and Moretton, 2007). Study carried out in Nepal found out that healthcare liquid wastes were loaded with multiple drug resistance bacteria and seemed to pose a huge public health threat in the transfer of such resistance to the bacterial pathogens causing community acquired infections, thereby limiting our antibiotic pool (Sharma *et al.*, 2010). It has been reported that in Sweden high prevalence of Vancomycin Resistant Enterococci (VRE) was demonstrated in Swedish sewage, which can be attributed to antimicrobial drugs or chemicals released into the sewage system that sustained VRE in the system (Aina *et al.*, 2002).

In Gaza, Al Shifa hospital wastewater was compared with non-health institutional wastewater in term of the spread of antibiotic resistance with non-health institution. The study revealed that most frequently identified bacterium was *Pseudomonas sp.* (33.1%) followed by *E. coli* (30.5%), *Enterococcus sp.* (21.4%), *Klebsiella sp.* (10.4%) and *Proteus sp.* (4.5%) in the hospital wastewaters. It was established that there was high incidence of antibiotic resistance among both gram-negative and gram-positive resistant bacteria isolated. It has been documented that in Brazil extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* were detected in effluents and sludge of a hospital sewage treatment plant. This result indicated the hospital wastewater treatment plant did not demonstrate a satisfactory significant efficacy in removing pathogenic micro-organisms by allowing for the dissemination of multi-resistant bacteria into the environment. Further documentation revealed that in U.S (wastewater research division, municipal environmental research laboratory, U.S. Environmental Protection Agency) revealed effect of ultra-violet (UV) light disinfection on antibiotic resistant coliforms in wastewater effluents. The document showed that UV irradiation effectively disinfected the wastewater effluent (the percentage of the total surviv-

ing coliform population resistant to Tetracycline or Chloramphenicol was significantly higher than the percentage of the total coliform population resistant to those antibiotics before UV irradiation and the finding was attributed to the mechanism of R-factor mediated resistance to tetracycline, Mark, 1982).

In Belgium, comparisons between the antimicrobial tolerance of Oxytetracycline-resistant heterotrophic bacteria isolated from hospital sewage and freshwater fish farm water were made. The study showed that Oxytetracycline-resistant hospital heterotrophs displayed a higher frequency (84%) of Ampicillin (Amp) tolerance compared to the Oxytetracycline-resistant heterotrophs from the freshwater fish-farm site (22%). In Austria, evaluations of antibiotic resistance of *E. coli* in sewage and sludge were conducted. The study revealed that the highest resistance rates were established in *E. coli* strains of a sewage treatment plant which treated both municipal sewage and sewage from a Hospital. In Nigeria, similar investigations demonstrated that hospital wastewaters played significant roles in the influence on the qualities of the bacteriological and physiochemical parameters on the receiving surface water bodies (environment), which can be attributed to increased amount of organic matter and essential nutrients in hospital wastewater. It was reported that most of the hospital effluents with other sewages released to water bodies (lakes, rivers, ponds or streams) or ground water bodies (hand dug and shallow wells), which are served as media for transfer of resistant bacteria and their genes in aquatic systems. As these water bodies are the major sources of drinking water (directly or indirectly) for human and animal consumption, its pollution may contribute to the maintenance bacterial antibiotic resistance in the environment and even the spread of bacterial antibiotic resistance (Mesdaghinia *et al.*, 2009).

In France, a study was conducted to trace whether antibiotic-resistant *Pseudomonas aeruginosa* isolated from hospitalized patients recovered in the hospital effluents. The study of genotyped both clinical and wastewater isolated antibiotic-resistant was determined by using pulsed-field gel electrophoresis (PFGE). The study revealed that there was no common PFGE pattern in antibiotic-resistant *P. aeruginosa* from humans and wastewater. It was concluded that antibiotic resistance profile of wastewater isolated antibiotic-resistant was different from that of clinical isolates (Tumeo *et al.*, 2008).

Jara *et al.* (2007) presented electro-oxidation treatment of aqueous solutions containing the antibiotics Ofloxacin and Lincomycin with different anodes (Ti/Pt, DSA1 type, graphite and three-dimensional (3D) electrode made of a fixed bed of activated carbon pellets). It was reported that effectiveness of the treatment of wastewater containing pharmaceuticals was assessed based on antibiotics and bacteria removal, as well as the electro-oxidation mechanism. The electro-oxidation was found to follow first order kinetics at the anode surface when Na_2SO_4 was at a low concentration (0.02 Normality). At lower concentration of Na_2SO_4 Ofloxacin was efficiently oxidized at the rate of $50 \text{ mgcm}^{-2}\text{A}^{-1} \text{ h}^{-1}$ for the bi-dimensional Ti/Pt electrode. It was observed that Lincomycin was oxidized with slow overall kinetics, which can be attributed to difficult deprotonation. It was concluded that the three-dimensional electrode would be the most appropriate for continuous industrial-scale process of aqueous solutions containing the antibiotics Ofloxacin and Lincomycin treatment.

In Watkinson *et al.* (2007), antibiotic resistance (AR) patterns of 462 *Escherichia coli* isolates from various wastewater, different surface water sources, and oysters were evaluated. The study revealed that the rates of AR and multiple-AR among isolates from surface water sources adjacent to wastewater treatment plant discharge sources were significantly higher ($p < 0.05$) than other *Escherichia coli* isolated from other sources. The rates of AR among *Escherichia coli* isolated from oysters exposed to wastewater treatment plant discharges were lower (less than 10%) than other *Escherichia coli* isolated from other sources.

Pedro *et al.* (2009) concluded that detrimental effects of anti-infective on aquatic microbiota are possible with the constant exposure of sensitive species. Indirect impact on human health cannot be ruled

out when considering the potential contribution of high anti-infective concentrations to the spreading of anti-infective resistance in bacteria.

Jury *et al.* (2010) concluded that precise explanations would be required to confirm these phenomena and experimental approaches, including culture-independent methods are suggested.

Emad and Malay (2010) concluded that the photo-Fenton process can be used as pre-treatment for antibiotic degradation and biodegradability enhancement of antibiotic wastewater containing amoxicillin and cloxacillin.

In Chelliapan *et al.* (2006; 2011) reported that anaerobic technology is a promising alternative for pharmaceutical wastewater treatment. In Chelliapan and Golar (2011), an Up-Flow Anaerobic Packed Bed reactor was employed to treat pharmaceutical wastewater containing antibiotics. Effect of organic loading rate (OLR) was assessed by modifying feed substrate concentration and hydraulic retention time (HRT). Performance of the system was based on COD removal, volatile fatty acid (VFA), gas production, methane yield and pH. The study showed that COD reduction was between 60 and 70% at an OLR between 1.5 and 4.6 kg COD.m⁻³.d⁻¹, which indicates that the biomass had acclimated to the antibiotics. An increase in the OLR soluble COD removal efficiency decreased to 50% at an OLR of 5.6 kg COD.m⁻³.d⁻¹, which indicates that OLR have effect on the methanogens.

Honarbakhsh (2011) examined the impact of antibiotic resistant bacteria from final sewages on water receiving bodies. On wastewater collected Phylogenetic analysis of the 16S rRNA gene sequences was performed to develop the composition and structure of the four microbial communities. The distribution of β -lactamase genes across the four sites was also assessed. Clone library information and TRFLP profiles revealed that the four communities were all controlled by Betaproteobacteria, which constituted approximately two thirds of the final effluent community. The diversity and abundance of other bacterial phylotypes varied across sites. Bacterial diversity included 9 phylotypes in the final effluent, 15 and 13 phylotypes respectively in the downstream. Bacteria were isolated from the same samples using culture-dependent techniques, which allowed to link antibiotic resistance to particular bacterial species. Clone libraries were dominated by Betaproteobacteria, culturable isolates mainly belonged to the class of Gammaproteobacteria. Analysis of the isolates from the final effluent exhibited the presence of *Bacillus*, *Enterobacter*, *Acinetobacter*, and *Staphylococcus* strains, while samples collected downstream from the plant were categorized by species belonging to the genera *Brevibacterium*, *Chryseobacterium*, *Aeromonas* and *Delftia*. All the isolates were resistant to amoxicillin, and most displayed resistance to multiple antibiotics. Phylogenetic analysis revealed that the distribution of the *bla* TEM4 gene is divided to two main clusters. The first cluster contains sequences that are exclusively detected at the plant and receiving water, but not at the upstream site, while sequences of the second cluster were present at all of the four sites.

Shreeshivadasan and Paul (2011) concluded that changes in the organic loading of pharmaceutical wastewater affected the treatment performance of the anaerobic reactor, and at high OLRs, anaerobic reactor was not able to withstand the short HRT, probably due to the intricacy of pharmaceutical wastewater.

Thompson *et al.* (2013) investigated the presence of methicillin-resistant *Staphylococcus aureus* (MRSA) in untreated hospital wastewaters (UHW), their transmission into the receiving sewage treatment plant (STP) and survival through the STP treatment. The study isolated 224 *Staphylococcus aureus* strains from UHW-1, UHW-2 and its receiving STP inlet (SI) and post-treatment outlet (SO). These strains were categorized using the PhP and RAPD-PCR methods. The strains were tested for their antibiotic resistance patterns. Resistance of the microbials to cefoxitin and the presence of *mecA* gene were used to identify MRSA isolates. In the study, 11 common (C) isolates and 156 single (S) PhP-RAPD isolates were identified, with two multidrug resistant (MDR) C- isolates found in H₂, SI and SO. These C- iso-

lates strains showed resistance to cefoxitin and vancomycin. The mean number of antibiotics to which the strains from UHWW were resistant (5.14 ± 2) was significantly higher than the STP isolates (2.9 ± 1.9 , $p < 0.0001$). Out of the total samples, 131 (68%) isolates were MRSA strains, 24 isolates were vancomycin resistant. MDR strains (including MRSA) were identified to be more prevalent in hospital wastewaters than in the STP. The study provided evidence of the survival of MRSA strains in UHWWs and their transit to the STP. It was reported that MRSA strains survived in the final treated effluent and chlorination stage. It was concluded that there is the needs to further investigate the load of MRSA in hospitals' wastewaters and possible their survival in STPs.

Katouli *et al.* (2012) evaluated the presence and survival of antibiotic resistant bacteria in untreated hospital wastewaters (UHWW) and transmission of antibiotic resistant bacteria in untreated hospital wastewaters to the receiving sewage treatment plant (STP) in South East Queensland (SEQ), Australia. The study revealed that in the *E. coli* strains collected from UHWW, seven common (C) PhP-RAPD species were found frequently and in most of weekly samples with multi-drug resistant (MDR) C-species found in STP-I. It was revealed that on three occasions multi-drug resistant (MDR) C-species found in STP-O. In *S. aureus* strains, seven C- species were frequently found in a majority of samples with two MDR C-species in STP-I. It was showed that on two occasions two MDR C-species in STP-O. The study concluded that some MDR bacterial strains found in UHWW may have the ability to survive, transmit to the STP and to the final treated effluent before being released into surface water bodies. Abargues *et al.* (2013) evaluated the effect of light, oxygen and microalgae on removal of the endocrine disrupting compounds 4-(1,1,3,3-tetramethylbutyl)phenol, technical-nonylphenol, 4-n-nonylphenol, and BPA from the effluent of an anaerobic MBR. Results demonstrated that the removal rates of these compounds differed with each post-treatment method and that aeration (oxygen) resulted in the highest removal ratios indicating the importance of an aerobic step in EDC degradation; a mass balance also showed that chemical oxidation was more important than sorption processes.

Choe *et al.* (2013) compared perchlorate-selective ion exchange, biological reduction and catalytic treatment for perchlorate removal based on the environmental impacts of consumable inputs normalized to the mass treated, in a lifecycle analysis. Because resource consumption during the operational phase comprises greater than 80% of the total impacts of these technologies, the analysis indicates that ion exchange is competitive, but emerging technologies also show great promise from an environmental sustainability perspective.

Kovalova *et al.* (2013) compared the removal efficiencies of micro-pollutants in hospital wastewater using O_3 , O_3/H_2O_2 , PAC, and UV light with or without TiO_2 . They found that the majority of wastewater micropollutants can be removed by ozone and PAC, while some recalcitrant compounds, like diatrizoate, can only be removed by high UV doses. Combined treatments of O_3/H_2O_2 and UV/ TiO_2 did not improve removal efficiency over single treatments of ozone and UV. The efficiency of MBRs and ozonation in removing endocrine disruptive potentials was studied for use in treatment of hospital wastewater (Maletz *et al.*, 2013). Treatment by MBRs reduced estrogenicity of sewage while ozonation efficiently removed most measured compounds. Yet another subset of studies in this category evaluated whole plants and either compared different treatment trains or compared processes within a single plant to each other in terms of effectiveness for emerging pollutant abatement. The concentrations and removal rates of ARGs were measured in three municipal WWTPs employing conventional treatment followed by different advanced treatment systems [biological aerated filter, constructed wetland, and UV disinfection] (Chen and Zhang 2013). The authors reported ARG reductions of 1 – 3 orders of magnitude through three conventional activated sludge processes. However, advanced post-treatment systems differed with orders

1–3 of reductions in ARGs in constructed wetlands, 0.6–1.2 orders in the biological aerated filter, and no apparent decrease by UV disinfection.

Flores *et al.* (2013) evaluated a full-scale drinking water treatment plant and several pilot treatment systems with a number of conventional (prechlorination, coagulation, sand filtration and ozonation) and advanced (RO, reverse electrodialysis and GAC adsorption) water treatment processes for their ability to remove perfluorinated compounds. Authors found the most efficient process was RO, which resulted in 99% removal of the studied pollutants. Other processes that showed removal of both the pollutants were adsorption onto GAC and reverse electrodialysis. All of the conventional processes proved to be fully ineffective. Concentrations of antibacterial chemicals triclocarban and triclosan and a triclosan biodegradation product methyltriclosan (MeTCS) were monitored in the liquid phase throughout a wastewater plant by Lozano *et al.* (2013). Triclocarban and triclosan were primarily removed by sorption and settling, though concentrations of both also decreased during nitrification denitrification and triclosan concentration decreased during the activated-sludge stage. Ozonation followed by sand filtration and PAC followed by either sand filtration or UF were tested in parallel for removal of 70 potentially problematic substances and reduction of ecotoxicological effects at large scale WWTPs over a year (Margot *et al.*, 2013).

Adina *et al.* (2013) reported that the degradation of 4-chlorophenol by an electrochemical advanced oxidation process. It was stated that several authors have proposed the following possible pathways of oxidation of some emerging pollutants by electrochemical processes. The pathways are as follows:

- i. 4-chlorophenol dechlorination to phenol;
- ii. hydroxylation of phenol to hydroquinone;
- iii. dehydrogenation of hydroquinone to benzoquinone;
- iv. oxidation of benzoquinone (with aromatic ring cleavage) to aliphatic carboxylic acids such as maleic acid, fumaric acid, malonic acid;
- v. oxidation of maleic and fumaric acids to oxalic acid,
- vi. formic acid to carbon dioxide and water

Rahim *et al.* (2014) revealed that the mean concentrations of GM, CHL and ceftazidime resistant bacteria in raw wastewater samples were 1.24×10^7 , 3.29×10^7 and 5.54×10^7 colony forming unit per 100 ml, respectively. There was a variation in prevalence of diverse groups of ARB in MWs and HWs. All WWTPs, there were decreased in the concentrations of ARB. It was established that high concentrations of ARB were found in the final effluent of WWTPs. In line with ARB, different groups of ARGs were found frequently in both MWs and HWs.

Sourav and Alok (2014) reviewed few treatment processes that are effective in remediating the pollution load in the pharmaceutical wastewater in an environmental sustainable manner (Table 2).

In Yamina *et al.* (2014), twelve cadmium-resistant bacteria were identified and isolated from hospital wastewaters. The cadmium-resistant isolated were characterized into both Gram-negative [75%] and Gram-positive [25%]. All isolated bacteria showed co-resistance to zinc and mercury in aqueous solutions. Isolated bacteria were antibiotic resistant and tolerated NaCl to a significant concentration of NaCl. The cadmium-resistant species (*Klebsiella pneumoniae* and *Pseudomonas aeruginosa*) were the most resistant to cadmium. The species *Klebsiella pneumonia* resisted to the highest concentrations of zinc and mercury. The results of toxicity tests on *Vibrio fischeri* revealed that the DI_{50} [15 minutes] as low as 0.07 carried away 50% luminescence inhibition.

Amenu (2014) revealed that there was no correlation between antibiotic-susceptibility and susceptibility to disinfectants in hospital strains and that there was cross-resistance between triclosan and antibiotics in *Pseudomonas aeruginosa*, which in turn showed that exposure of a clinically significant bacterium to the antiseptic triclosan efficiently can select for multi-drug resistant (MDR) derivatives, which include high-level resistance to an ant *pseudomonas* drug.

Rongjun *et al.* (2015) conclude that RBCs are effective in the treatment of wastewater, but RBCs are rarely use for treating antibiotic pharmaceutical wastewater.

Benson *et al.* (2015) revealed that there was a very high phenotypic multiple resistance within the isolates, which were confirmed by genetic analyses.

Birošová, *et al.* (2015) concluded that stabilized sludges represent sources of methicillin, penicillin, ampicillin and ciprofloxacin resistant *S. aureus*.

Huan *et al.*, (2018) shows that the emission of dissolved organic matter in wastewater after electrochemical treatment could accelerate estrogen degradation and play a positive role on the pollutant transformation in the environment.

Electrochemical Treatment of Simulated Pharmaceutical Wastewater

Tables 3 and 4 present summary and results of statistical analysis (fractional factorial experiment and analysis of variance) of the effect of the selected factors on the performance of electrochemical treatment of pharmaceutical wastewaters. Table 3 revealed that performance of the treatment can be classified into two categories as performance greater than 54.88% and performance less than 54.88%. Experiments 2, 11, 12, 15 and 16 provided performance higher 54.88% which indicated that factors at higher levels in the above stated experiments had positive effects on the performance of electrochemical treatment of pharmaceutical wastewaters. From the tables, 54.88% was the mean performance of the process (expected performance when all the selected factors are at their low levels). Experiments 3,4,5,6, 7, 8, 9, 10, 13 and 14 were performance less than 54.88%. This result shows that factors at high level in experiments 3,4,5,6, 7, 9, 10, 13 and 14 contributed negatively (reduced) to the performance of electrochemical treatment of pharmaceutical wastewaters.

For effective performance of electrochemical system the results indicated that experiments 3,4,5,6, 7, 8, 9, 10, 13 and 14 needs to be minimized, while experiments 2, 11, 12, 15 and 16 must be optimized to save operational cost. Tables 4(a and b) present the detailed statistical analysis. Table 4a presents summary of analysis of variance and Table 3b shows basic computation of the analysis. From these Tables factors with significant effect were current applied ($F_{1,7} = 9.80$), concentration of $\text{Ca}(\text{OCl})_2$ ($F_{1,7} = 11.34$) and volume of the wastewater used ($F_{1,7} = 33.79$) at 95% confidence level ($F_{1,7} = 5.59$ critical value). Separation distance between the electrodes ($F_{1,7} = 4.27$) was another factor that had significant effect at 90% confidence level. Although, other factors (Concentration of BOD_5 , Depth of the Electrodes in the Wastewater, Temperature of the Wastewater and Treatment Time) had effect on the performance of the treatment process, but the effects were not significant at 90% confidence level. These can be attributed to the presence of $\text{Ca}(\text{OCl})_2$, which acted as oxidizing agent and electrolyte to speed up the treatment process.

Electrochemical and Chemical Treatment of Typical Pharmaceutical Wastewater

Table 5 shows the results of performance evaluation of each the selected wastewater treatment methods in respect to wastewaters from pharmaceutical industry. From the table efficacies of electrochemical

method in removing chromium, BOD₅, SS, colour, chloride, total solids, lead, cadmium and zinc were 100, 88.8, 96.6, 80.0, 12.5, 92.9, 100, 100 and 76.7% using 0.43 MJ (30Wh per litre). For chemical treatment method 82.5, 86.0, 70.2, 60, -100, 48.2, 100, 100, and 65.6% efficacies were achieved using 1200 mg/L of Ca(OCl)₂. For modified electrochemical treatment method 100, 99.1, 98.6, 90.0, -205, 58.2, 100, 100 and 80.5% efficacies were achieved using 1.61MJ(112Wh per litre) and for combination treatment of electrochemical and chemical treatment method efficacies of 100, 99.1, 99.4, 80.0, -20, 92.9, 100, 100 and 85.4% were attained using 0.43 MJ (30 Wh per litre) and 450 mg/L of Ca(OCl)₂. Efficacies variation by electrochemical treatment may be attributed to quantity of current through the wastewater. Comparing quality of treated wastewaters with environmental standard limits in Nigeria, it can be stated that with exception of BOD and suspended solids concentration other pollutants met FEPA limits, indicating the method can be used effective for pharmaceutical wastewaters but must be followed by sedimentation process to remove suspended solids and left over BOD.

Cost analyses: Comparison of the performance and efficiency of the three technologies studied can be summarized as follows: under similar conditions of aluminium concentration and pH the three technologies attain similar results. For this reason, comparison of the operating costs of the three processes is a critical issue in recommending the use of one of the technologies. In this context, it is important to note that the aim of this study has been to compare the operating costs of the three dosing technologies, and to estimate the total costs of each one. Thus, a preliminary economical evaluation has been conducted, in which the items that can be considered to be the same in both processes, such as labour cost and management of the sludge have not been included in the calculations. In this context, the following assumptions have been made for the economical estimation (Pablo et al., 2008):

- The labour costs required in the three processes should be similar, despite the management of reagents involved in the conventional dosing of reagents, because the three processes can easily be automated. Also, the amount of sludge produced in the three processes must be similar, and therefore, the costs related to management of the sludge can be considered to be the same. The technique of cost estimate in literature such as Oladepo *et al.* (2011) was adopted for capital and operational costs.
- The electrocoagulation cell is properly designed, that is, the turbulence generated by the gas bubbles emerging from the electrode surfaces is conveniently used to promote mixing of the electrodisolved coagulants, and the electroflocculation process. In principle, this fact should not increase the cost of the cell, as a simple layout of the electrodes in a parallel and slightly slanting position allows promotion of both effects. As a consequence, additional mechanical mixing is not required to promote either coagulation or flocculation processes. In addition, in the electrochemical process it is not necessary to consider the energy consumption derived from the dissolution of solid coagulants (salts) to prepare the liquid reagent to dose the coagulation process.
 - To estimate the power necessary for mixing in conventional coagulation and modified conventional coagulation processes, the velocity gradient G (s⁻¹) has been used. This parameter is a function of the mixing velocity, the type of impeller, and the ratio between the tank and the impeller sizes as follows:

$$G = \sqrt{\frac{P}{\mu \times V}} \quad (9)$$

Typical values of parameter G are reported in the literature for the coagulation stage: 900 s^{-1} for 30 s, and for the flocculation stage: 25 s^{-1} for 30 minutes. To calculate the mixing power necessary to prepare the coagulant solution, the author used a medium value for G of 700 s^{-1} for 10 minutes (Viessman and Hammer, 1993; Pablo *et al.*, 2008).

- The operating costs of the conventional and modified conventional coagulation process involve the costs related to the price of the coagulant reagents, and the energy consumption required for mixing in coagulation and flocculation processes. In addition, solid coagulants are used, it is necessary to consider the mixing required to prepare the coagulant solution. The energy consumption related to the pumping of the coagulant solutions was not considered to be negligible. To estimate the energy costs, one must take into account that the prices of electricity are highly dependent on the particular country. The present unitary electricity cost for industrial use in Nigeria is around $0.23 \$\text{kWh}^{-1}$ (currently $1 \text{ US\$} = 365 \text{ Nigeria Naira}$).
 - The operating costs of the electrocoagulation process involve the price of the aluminium rod (used as electrodes) and the energy costs for their electrodisolution. The results obtained in the study of electrodisolution of aluminium have been borne in mind in estimating the energy consumption for the electrodisolution of aluminium. Previous study has shown that the concentrations of aluminium electro-dissolved are more than those values predicted by Faraday's Law, due to chemical corrosion of the electrodes. Thus, the aluminium concentrations have been related to the Faraday's Law by an empirical equation as follows (Pablo *et al.*, 2008):

$$C_{Al} = 100 \left(\left(\frac{I \times W_{Al}}{F \times n \times q} \right) X \right) \quad (10)$$

In electrochemical process the power consumption P (W) is given as follows:

$$C_E = \left(\frac{I \times E_v}{C_{Al} \times q} \right) = 100 \left(\frac{F \times n \times E_v}{W_{Al} \times X} \right) \quad (11)$$

The voltage (E_v) in the electrodisolution of aluminium has been set for this estimation according to the experimental results obtained in previous work, although this parameter is strongly dependent on the conductivity of the raw water treated.

- Regarding the separation of solids (pollutants with reduced solubility), in the case of conventional coagulation, the most favourable conditions have been assumed: the solids can settle rapidly, and therefore, it is not necessary to use the dissolved air flotation processes that would increase the cost of the treatment. In this context, in the electrochemical process the gas bubbles produced on the surface of the electrodes promote the separation of solids to the top of the cell (electroflotation process). Taking all of this into account, the separation of solids has not been considered to increase the cost of either of the coagulation processes.

Concerning the operational comparison of the three processes, it was found that the changes in pH are different in the three dosing technologies. For this reason, it was necessary to adjust the final pH of the raw wastewater to obtain recommended pH value. Figure 6a presents relationship between cost of treatment and efficacy of selected wastewater treatment techniques in removing

selected pollutants from pharmaceutical wastewater. Figure 6b shows relationship between rate of change in cost of treatment per unit percentage and selected treatment techniques used. From the figures it can be seen that electrochemical and modified electrochemical (assisted electrochemical) treatment techniques are among the cost effective techniques in treating pharmaceutical wastewater.

FUTURE RESEARCH DIRECTIONS

As a consequence of these uncertainties and increase in diseases, the designs as well as management strategies of water and wastewater treatment plants have to include effects of EPs on biological treatment processes, expected end products after chemical oxidation of EPs and management of wastes from treatment plant.

CONCLUSION

It can be concluded based on the study that:

- a) wastewaters from pharmaceutical activities contain emerging pollutants,
- b) there are more than 1036 emerging pollutants and various concept for management of emerging pollutants in the aquatic environment,
- c) there are twenty-one antibiotic-multi-resistant strains,
- d) in Belgium, Oxytetracycline-resistant hospital heterotrophs displayed a higher frequency (84%) of Ampicillin (Amp) tolerance compared to the Oxytetracycline-resistant heterotrophs from the freshwater fish-farm site,
- e) in Austria, the highest resistance rates were established in *E. coli* strains of a sewage treatment plant,
- f) in Nigeria, most of the hospital effluents with other sewages released to water bodies are served as media for transfer of resistant bacteria and their genes in aquatic systems,
- g) in France, antibiotic resistance profile of wastewater isolated antibiotic-resistant was different from that of clinical isolates and there are cadmium-resistant species (*Klebsiella pneumoniae* and *Pseudomonas aeruginosa*),
- h) in Gaza, most frequently identified bacterium was *Pseudomonas sp.* (33.1%) followed by *E. coli* (30.5%), *Enterococcus sp.* (21.4%), *Klebsiella sp.* (10.4%) and *Proteus sp.* (4.5%) in the hospital wastewaters,
- i) In France, there was no common pulsed-field gel electrophoresis pattern in antibiotic-resistant *P. aeruginosa* from humans and wastewater,
- j) In hospital wastewater the dominated antibiotic resistant bacteria were ciprofloxacin, penicillin, methicillin, ampicillin and vancomycin resistants,
- k) the bacterial population resistant to disinfectants were *Enterobacteriaceae*, *Staphylococcus spp*, and *Bacillus spp*, which are highly associated to nosocomial infections,
- l) current applied through the electrodes, concentration of $\text{Ca}(\text{OCl})_2$, volume of the wastewater and separation distance between the electrodes were significant factors that influence electrochemical treatment of pharmaceutical wastewaters at 90% confidence level,
- m) Electrochemical treatment of pharmaceutical wastewaters using public electricity source is economical when compared with other treatment processes evaluated.

REFERENCES

- Aboulhassan, M. A., Souabi, S., Yaacoubi, A., & Baudu, M. (2006). Removal of surfactant from industrial wastewaters by coagulation flocculation process. *International Journal of Environmental Science and Technology*, 3(4), 327–332. doi:10.1007/BF03325941
- Adams, C., Asce, M., Wang, Y., Loftin, K., & Meyer, M. (2002). Removal of antibiotics from surface and distilled water in conventional water treatment processes. *Journal of Environmental Engineering*, 128(3), 253–260. doi:10.1061/(ASCE)0733-9372(2002)128:3(253)
- Adina, E. S., Cristina, O., Carmen, L., Paula, S., Paulina, V., Cornelia, B., & Ioan, G. (2013). Waste Water Treatment Methods. doi:10.5772/53755
- Aina, I., Inger, K., Anders, F., & Roland, M. (2002). High Prevalence of Vancomycin-Resistant Enterococci in Swedish Sewage. *ASM. Applied and Environmental Microbiology*, 2838–2842.
- Alder, A. C., McArdell, C. S., Golet, E. M., Molnar, E. M., Kuch, H. M., & Giger, W. (2003). Environmental exposure of antibacterial agents in hospital and municipal wastewater and river water in Switzerland [Abs.]. *Proceedings of 3rd International Conference on Pharmaceuticals and Endocrine Disrupting Chemicals in Water*, 236–237.
- Amenu, D. (2014). Assessment of Disinfectant and Antibiotic Resistant Bacteria in Hospital Wastewater. *World Journal of Life Sciences Research*, 1(1), 1–7.
- American Public Health Association (APHA). (1998). Standard methods for examination of water and wastewater. American Public Health Association, American Water Works Association and Water Pollution Control Federation.
- Andreozzi, R., Canterino, M., Marotta, R., & Paxeus, N. (2005). Antibiotic removal from wastewaters: The ozonation of amoxicillin. *Journal of Hazardous Materials*, 122(3), 243–250. doi:10.1016/j.jhazmat.2005.03.004 PubMed
- Benson, C. I., Sisipho, G., Larry, C. O., Uchechukwu, U. N., & Anthony, I. O. (2015). Antibiotic Susceptibilities of Enterococcus Species Isolated from Hospital and Domestic Wastewater Effluents in Alice, Eastern Cape Province of South Africa. *International Journal of Environmental Research and Public Health*, 12(4), 4231–4246. doi:10.3390/ijerph120404231 PubMed
- Berge, A. C. B., Dueger, E. L., & Sisco, W. M. (2006). Comparison of Salmonella enterica serovar distribution and antibiotic resistance patterns in wastewater at municipal water treatment plants in two California cities. *Journal of Applied Microbiology*.
- Best, M., Kennedy, M. E., & Coates, F. (1990). Efficacy of a Variety of Disinfectants against Listeria spp. *ASM. Applied and Environmental Microbiology*, 2(2), 377–380. doi:10.1128/AEM.56.2.377-380.1990 PubMed
- Birošová, L., Nagyová, K., Mackuřák, T., & Medvedřová, A. (2015). The Occurrence of Antibiotic Resistant Escherichia Coli and Staphylococcus Aureus In Wastewater and Sludge From Treatment Plant In Czech Republic. *Proceedings of the 14th International Conference on Environmental Science and Technology*.

Chang, X., Meyer, M. T., Liu, X., Zhao, Q., Chen, H., Chen, J., Qiu, Z., Yang, L., Cao, J., & Shu, W. (2010). Determination of antibiotics in sewage from hospitals, nursery and slaughter house, wastewater treatment plant and source water in Chongqing region of Three Gorge Reservoir in China. *Environ. Pollut.*, 158(5), 1444–1450. doi:10.1016/j.envpol.2009.12.034 PubMed

Chelliapan, S., & Golar, S. (2011). Pre-treatment of Antibiotic Wastewater Using an Anaerobic Reactor. *Research Journal of Chemical Sciences*, 1(6), 59–63.

Chelliapan, S., Wilby, T., & Sallis, P. J. (2006). Performance of an up-flow anaerobic stage reactor (UASR) in the treatment of pharmaceutical wastewater containing macrolide antibiotics. *Water Research*, 40(3), 507–516. doi:10.1016/j.watres.2005.11.020 PubMed

Chen, G. (2004). Electrochemical technologies in wastewater treatment. *Separation and Purification Technology*, 38(1), 11–41. doi:10.1016/j.seppur.2003.10.006

Chen, H., & Zhang, M. (2013). Effects of Advanced Treatment Systems on the Removal of Antibiotic Resistance Genes in Wastewater Treatment Plants from Hangzhou, China. *Environmental Science & Technology*, 47(15), 8157–8163. PubMed

Choi, K.-J., Son, H.-J., & Kim, S.-H. (2007). Ionic treatment for removal of sulphonamide and tetracycline classes of antibiotic. *The Science of the Total Environment*, 387(1-3), 247–256. doi:10.1016/j.scitotenv.2007.07.024 PubMed

Collado, N., Buttiglieri, G., Marti, E., Ferrando-Climent, L., Rodriguez-Mozaz, S., Barcelo, D., Comas, J., & Rodriguez-Roda, I. (2013). Effects on Activated Sludge Bacterial Community Exposed to Sulfamethoxazole. *Chemosphere*, 93(1), 99–106. doi:10.1016/j.chemosphere.2013.04.094 PubMed

Davore, J. L., & Farnum, N. R. (1999). *Applied Statistics for Engineers and Scientists* (1st ed.). Duxbury Press.

Deniz, N., Angela, R., Dimitrios, B., & Viviane, Y. (2012). Removal of the antibiotic levofloxacin (LEVO) in water by ozonation and TiO₂ photocatalysis. *Chemical Engineering Journal*, 189–190, 41–48.

Devore, J. L. (2000). *Introduction to Probability and Statistics for Engineers and Scientists* (1st ed.). Duxbury Thomas Learning.

Doolette, C., McLaughlin, M., Kirby, J., Batstone, D., Harris, H., Ge, H., & Cornelis, G. (2013). Transformation of PVP Coated Silver Nanoparticles in a Simulated Wastewater Treatment Process and the Effect on Microbial Communities. *Chemistry Central Journal*, 7(1), 1–18. doi:10.1186/1752-153X-7-46 PubMed

Dubreuil, L., Devos, J., Neut, C., & Romond, C. (1984). Susceptibility of Anaerobic Bacteria from Several French Hospitals to Three Major Antibiotics. *ASM. Antimicrobial Agents and Chemotherapy*, 25(6), 764–766. doi:10.1128/AAC.25.6.764 PubMed

Elmolla, E., & Chaudhuri, M. (2009a). Improvement of biodegradability of synthetic amoxicillin wastewater by photo-Fenton process. *World Applied Sciences Journal*, 5, 53–58.

Elmolla, E., & Chaudhuri, M. (2009b). Optimization of Fenton process for treatment of amoxicillin, ampicillin and cloxacillin antibiotics in aqueous solution. *Journal of Hazardous Materials*, 170(2-3), 666–672. doi:10.1016/j.jhazmat.2009.05.013 PubMed

Emad, S. E., & Malay, C. (2010). Photo-Fenton Treatment of Antibiotic Wastewater. *Nature Environment and Pollution. An International Quarterly Scientific Journal Technology.*, 9(2), 365–370.

Flores, C., Ventura, F., Martin-Alonso, J., & Caixach, J. (2013). Occurrence of Perfluorooctane Sulfonate (PFOS) and Perfluorooctanoate (PFOA) in N.E. Spanish Surface Waters and Their Removal in a Drinking Water Treatment Plant That Combines Conventional and Advanced Treatments in Parallel Lines. *The Science of the Total Environment*, 461–462, 618–626. doi:10.1016/j.scitotenv.2013.05.026 PubMed

Frank, W., Adair, S., Geftic, G., & Justus, G. (1969). Resistance of *Pseudomonas* to Quaternary Ammonium Compounds. *ASM. Applied Microbiology*, 299–302.

Gardiner, W. P., & Gettinby, G. (1998). *Experimental Design Techniques in Statistical Practice* (2nd ed.). Harwood Publishing., doi:10.1533/9780857099785.

Geissena, V., Hans, M., Erwin, K., Günter, U., & Marti, N. (2015). Emerging pollutants in the environment: A challenge for water resource management. *International Soil and Water Conservation Research*, 3(1), 57–65. doi:10.1016/j.iswcr.2015.03.002

Gerald, M., & Russell, A. D. (1999). Antiseptics and Disinfectants: Activity, Action, and Resistance. *ASM. Clinical Microbiology Reviews*, 12(1), 147–179. doi:10.1128/CMR.12.1.147 PubMed

Ghauch, A., Tuqan, A., & Assi, H. A. (2009). Antibiotic removal from water: Elimination of amoxicillin and ampicillin by microscale and nanoscale iron particles. *Environ. Pollution.*, 157(5), 1626–1635. doi:10.1016/j.envpol.2008.12.024 PubMed

Golet, E. M., Alder, A. C., & Giger, W. (2002). Environmental exposure and risk assessment of fluoroquinolone antibacterial agents in wastewater and river water of the Glatt Valley Watershed. Switzerland. *Environmental Science & Technology*, 36(17), 3645–3651. doi:10.1021/es0256212 PubMed

Golet, E. M., Xifra, I., Siegrist, H., Alder, A. C., & Giger, W. (2003). Environmental exposure assessment of fluoroquinolone antibacterial agents from sewage to soil. *Environmental Science & Technology*, 37(15), 3243–3249. doi:10.1021/es0264448 PubMed

Guttman, I., Wilks, S. S., & Hunter, J. S. (1971). *Introductory Engineering Statistics* (2nd ed.). John Wiley and Sons Inc.

Hirsch, R., Ternes, T., Haberer, K., & Kratz, K. L. (1999). Occurrence of antibiotics in the aquatic environment. *The Science of the Total Environment*, 225(1-2), 109–118. doi:10.1016/S0048-9697(98)00337-4 PubMed

Honarbaksh, M. (2011). Impact of Wastewater Effluent on the Structure and Antibiotic Resistance of Native Microbial Communities. A Master of Science Graduate Program in Microbiology and Molecular Genetics thesis submitted to the Graduate School-New Brunswick Rutgers, The State University of New Jersey and The Graduate School of Biomedical Sciences University of Medicine and Dentistry.

Hong, A. D., Ngoc, H. P., Hoang, T. N., Thi, T. H., Hung, V. P., Van, C. P., Michael, B., Walter, G., & Alfredo, C. A. (2008). Occurrence, fate and antibiotic resistance of fluoroquinolone antibacterials in hospital wastewaters in Hanoi, Vietnam. *Chemosphere*, 72(6), 968–973. doi:10.1016/j.chemosphere.2008.03.009 PubMed

Ivan, S., Marin, M., Helena, K., Jakopovic, S. T., Josip, C., Ivan, M., & Marijan, A. (2011). Removal of antimicrobials using advanced wastewater treatment. *Journal of Hazardous Materials*, 192, 319–328. PubMed

Jara, C. C., Fino, D., Specchia, V., Saracco, G., & Spinelli, P. (2007). Electrochemical removal of antibiotics from wastewaters. *Applied Catalysis B: Environmental*, 70(1-4), 479–487. doi:10.1016/j.apcatb.2005.11.035

Jury, K.L. Vancov, T, Stuetz, R.M., & Khan, S.J. (2010). Antibiotic resistance dissemination and sewage treatment plants. Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology, 509 – 519.

Katouli, M., Thompson, J.M., Gündoğdu, A., & Stratton, H.M. (2012). Antibiotic Resistant Bacteria in Hospital Wastewaters and Sewage Treatment Plants. Science Forum and Stakeholder Engagement: Building Linkages. Collaboration and Science Quality, 225-229.

Kosma, C. I., Lambropoulou, D. A., & Albanis, T. A. (2010). Occurrence and removal of PPCPs in municipal and hospital wastewaters in Greece. *Journal of Hazardous Materials*, 179(1-3), 804–817. doi:10.1016/j.jhazmat.2010.03.075 PubMed

Kosutic, K., Dolar, D., Asperger, D., & Kunst, B. (2007). Removal of antibiotic from model wastewater by RO/NF membrane. *Separation and Purification Technology*, 53(3), 244–249. doi:10.1016/j.seppur.2006.07.015

Kovalova, L., Siegrist, H., Von Gunten, U., Eugster, J., Hagenbuch, M., Wittmer, A., Moser, R., & Mc Ardell, C. S. (2013). Elimination of Micropollutants During Post- Treatment of Hospital Wastewater with Powdered Activated Carbon, Ozone, and UV. *Environmental Science & Technology*, 47(14), 7899–7908. doi:10.1021/es400708w PubMed

Li, H., Zhiwei, S., & Changxiong, Z. (2014). Trends in research on electrochemical oxidation. *Croatica Chemica Acta*, 87(2), 185–194. doi:10.5562/cca2446

Lozano, N., Rice, C. P., Ramirez, M., & Torrents, A. (2013). Fate of Triclocarban, Triclosan and Methyltriclosan During Wastewater and Biosolids Treatment Processes. *Water Research*, 47(13), 4519–4527. doi:10.1016/j.watres.2013.05.015 PubMed

Maletz, S., Floehr, T., Beier, S., Klumper, C., Brouwer, A., Behnisch, P., Higley, E., Giesy, J. P., Hecker, M., Gebhardt, W., Linnemann, V., Pinnekamp, J., & Hollert, H. (2013). In Vitro Characterization of the Effectiveness of Enhanced Sewage Treatment Processes to Eliminate Endocrine Activity of Hospital Effluents. *Water Research*, 47(4), 1545–1557. doi:10.1016/j.watres.2012.12.008 PubMed

Manisankar, P., Rani, C., & Viswanathan, S. (2004). Effect of halides in the electrochemical treatment of distillery effluent. *Chemosphere*, 57(8), 961–966. doi:10.1016/j.chemosphere.2004.07.026 PubMed

Margot, J., Kienle, C., Magnet, A., Weil, M., Rossi, L., de Alencastro, L. F., Abegglen, C., Thonney, D., Chevre, N., Scharer, M., & Barry, D. A. (2013). Treatment of Micropollutants in Municipal Wastewater: Ozone or Powdered Activated Carbon? *The Science of the Total Environment*, 461–462, 480–498. doi:10.1016/j.scitotenv.2013.05.034 PubMed

Mark, C. M. (1982). Effect of UV Light Disinfection on Antibiotic-Resistant Coliforms in Wastewater Effluents. *ASM. Applied and Environmental Microbiology*, 43(2), 371–377. doi:10.1128/AEM.43.2.371-377.1982 PubMed

Martin, S. F., & Charles, H. D. (1966). Factors Influencing the Occurrence of High Numbers of Iodine-Resistant Bacteria in iodinated Swimming Pools. *ASM. Applied Microbiology*, 14(4), 627–635. doi:10.1128/AEM.14.4.627-635.1966 PubMed

Melanie, V., Indra, B., Anna, N., & Martin, J. (2017). Comparison of UVA and UVC irradiation of FQs and their MICs for sustainable water treatment. 2nd Green and Sustainable Chemistry Conference at Berlin. https://www.researchgate.net/publication/316977230_Comparison_of_UVA_and_UVC_irradiation_of_FQs_and_their_MICs_for_sustainable_water_treatment?enrichId=rgreq-576f046106688384d9567cd-1fe917cea-XXX&enrichSource=Y292ZXJQYWdlOzMxNjk3NzIzMDtBUzo0OTQ3MTg3NDU3N-TU2NDhAMTQ5NDk2MTU0ODQyOQ%3D%3D&el=1_x_3&_esc=publicationCoverPdf. Accessed on 22nd March 2019

Mesdaghinia, A. R., Naddafi, K., Nabizadeh, R., Saeedi, R., & Zamanzadeh, M. (2009). Wastewater Characteristics and Appropriate Method for Wastewater Management in the Hospitals. *Iranian J Pub. Health*, 38(1), 34–40.

Michael, A. G., Mark, D. C., Kevin, B. A., & Jayne, C. B. (1985). Transfer of Plasmids pBR322 and pBR325 in Wastewater from Laboratory Strains of *Escherichia coli* to Bacteria Indigenous to the Waste Disposal System. *ASM. Applied and Environmental Microbiology*, 836–841.

Micheal, V. W., & John, W. V. (1985). Occurrence of Multiple-Antibiotic-Resistant Enteric Bacteria in Domestic Sewage and Oxidation Lagoons. *Applied and Environmental Microbiology*, 930–933. PubMed

Naumczyk, J., Szpyrkowicz, L., & Zilio-Grandi, F. (1996). Electrochemical treatment of textile wastewater. *Water Science and Technology*, 34(11), 17–24. doi:10.2166/wst.1996.0258

NORMAN. (2019). Network of reference laboratories, research centres and related organisations for Monitoring of emerging environmental substances. <https://www.norman-network.net/?q=Home>

Núñez, L., & Moreton, J. (2007). Disinfectant-resistant bacteria in Buenos Aires city hospital wastewater. *Brazilian Journal of Microbiology*, 38(4), 123–132. doi:10.1590/S1517-83822007000400012

Obst, U., Schwartz, T., & Volkmann, H. (2006). Antibiotic resistant pathogenic bacteria and their resistance genes in bacterial biofilms. *The International Journal of Artificial Organs*, 29(4), 387–394. doi:10.1177/039139880602900408 PubMed

Oke, I. A. (2009). Orthogonal Experiments in the Development of Carbon –Resin For Chloride Ions Removal. *Statistical Methodology*, 6(2), 109–119. doi:10.1016/j.stamet.2008.04.004

Oke, L. A., Oladepo, K. T., & Ogedengbe, M. O. (2007b). Utilization of 2k factorial experiments for the determination of factors that influence electrochemical process. *Ife Journal Technol.*, 19(2), 48–54.

Oke, L. A., Umoru, L. E., & Ogedengbe, M. O. (2007a). 2k Factorial experiments on factors that influence stability of carbon resin electrodes. *FUTAJEET*, 5(2), 135–141.

Oke, L. A., Umoru, L. E., & Ogedengbe, M. O. (2007c). Properties and stability of a carbon-resin electrode. *J. Mater. Design*, 28(7), 2251–2254. doi:10.1016/j.matdes.2006.05.019

Oladepo, K. T., Oke, I. A., & Ogedengbe, M. O. (2011). Evaluation of An Electrochemical Method of Coagulation for Surface Water Treatment. *Ife Journal of Technology, Special Issue.*, 20(2), 81–88.

Olowe, O. A., Olayemi, A. B., Eniola, K. I. T., & Adeyeba, O. A. (2004). Antibacterial activity of some selected disinfectants regularly used in hospitals. *African Journal of Clinical and Experimental Microbiology*, 5(1), 126–130. doi:10.4314/ajcem.v5i1.7365

Ötöker, H. M., & Akmehmet-Balcioglu, I. (2005). Adsorption and degradation of enrofloxacin, a veterinary antibiotic on natural zeolite. *Journal of Hazardous Materials*, 122(3), 251–258. doi:10.1016/j.jhazmat.2005.03.005 PubMed

Pablo, C. F., Mart'inez, C. J., Cristina, S., & Manuel, A. R. (2009). Technical and economic comparison of conventional and electrochemical coagulation processes. *Journal of Chemical Technology and Biotechnology (Oxford, Oxfordshire)*, 84(5), 702–710. doi:10.1002/jctb.2102

Panizza, M., & Cerisola, G. (2003). A comparative study on direct and indirect electrochemical oxidation of polyaromatic compounds. *Ann Chim.*, 93(12), 977–984. PubMed

Park, S., & Choi, K. (2008). Hazard assessment of commonly used agricultural antibiotics on aquatic ecosystems. *Ecotoxicology (London, England)*, 17(6), 526–538. doi:10.1007/s10646-008-0209-x PubMed

Pauwels, B., & Verstraete, W. (2006). The treatment of hospital wastewater: an appraisal The treatment of hospital wastewater: an appraisal. *Journal of Water and Health*, 04(4), 406–416. doi:10.2166/wh.2006.0024 PubMed

Pedro, A. S., Matthieu, F., Christian, G., & Sébastien, S. (2009). Review of the Occurrence of Anti-infectives in Contaminated Wastewaters and Natural and Drinking Waters. *Environmental Health Perspectives*, 117(5), 675–684. doi:10.1289/ehp.11776 PubMed

Pomati, F., Castiglioni, S., Zuccato, E., Fanelli, R., Vigetti, D., Rossetti, C., & Davide, C. (2006). Effects of a complex mixture of therapeutic drugs at environmental levels on human embryonic cells. *Environmental Science & Technology*, 40(7), 2442–2447. doi:10.1021/es051715a PubMed

Putra, E. K., Pranowo, R., Sunarso, J., Indraswati, N., & Ismadji, S. (2009). Performance of activated carbon and bentonite for adsorption of amoxicillin from wastewater: Mechanism, isotherms and kinetics. *Water Research*, 43(9), 2419–2430. doi:10.1016/j.watres.2009.02.039 PubMed

Rahim, A., Mahnaz, N., Hossein, K., & Akbar, H. (2014). Monitoring and Comparison of Antibiotic Resistant Bacteria and Their Resistance Genes in Municipal and Hospital Wastewaters. *International Journal of Preventive Medicine*, 5(7), 887–894. PubMed

Rana, P., Mohan, N., & Rajagopal, C. (2004). Electrochemical removal of chromium from wastewater by using carbon aerogel electrodes. *Water Research*, 38(12), 2811–2820. doi:10.1016/j.watres.2004.02.029 PubMed

Ravera, M., Ciccarelli, C., Gianotti, V., Scorza, S., & Osella, D. (2004). Electroassisted methods for waste destruction: Silver (II) and peroxydisulfate reagents in the electrochemically mediated oxidation of polyaromatic sulfonates. *Chemosphere*, 57(7), 587–594. doi:10.1016/j.chemosphere.2004.08.035 PubMed

Roberto, A., Marisa, C., Raffaele, M., & Nicklas, P. (2005). Antibiotic removal from wastewaters: The ozonation of amoxicillin. *Journal of Hazardous Materials*, 122(3), 243–250. doi:10.1016/j.jhazmat.2005.03.004 PubMed

Rodney, M. D., & William, J. C. B. (2002). Survival Mechanisms of Clinically Relevant Microorganisms. ASM. *Clinical Microbiology Reviews*, 15(2), 167–193. doi:10.1128/CMR.15.2.167-193.2002 PubMed

Saichang, Z., Hong, C., & Jianan, L. (2013). Sources, distribution and potential risks of pharmaceuticals and personal care products in Qingshan Lake basin, Eastern China. *Ecotoxicology and Environmental Safety*, 96, 154–159. doi:10.1016/j.ecoenv.2013.06.033 PubMed

Sánchez-Polo, M., Rivera-Utrilla, J., Prados-Joya, G., Ferro-García, M. A., & Bautista-Toledo, I. (2008). Removal of pharmaceutical compounds, nitroimidazoles, from waters by using the ozone/carbon systems. *Water Research*, 42(15), 4163–4171. doi:10.1016/j.watres.2008.05.034 PubMed

Sendzik, J., Shakibaei, M., Schafer-Korting, M., & Stahlmann, R. (2005). Fluoroquinolones cause changes in extracellular matrix, signalling proteins, metalloproteinases and caspase-3 in cultured human tendon cells. *Toxicology*, 212(1), 24–36. doi:10.1016/j.tox.2005.04.002 PubMed

Sharma, D. R., Pradhan, B., & Mishra, S. K. (2010). Multiple drug resistance in bacterial isolates from liquid wastes generated in central hospitals of Nepal. *Kathmandu University Medical Journal*, 8(1), 29, 40–44.

Shreeshivadasan, C., & Paul, J. S. (2011). Application of Anaerobic Biotechnology for Pharmaceutical Wastewater Treatment. *Special Issue on Environmental Management For Sustainable Development.*, 2(1), 13–21.

Sim, W. J., Lee, J. W., Lee, E. S., Shin, S. K., Hwang, S. R., & Oh, J. E. (2011). Occurrence and distribution of pharmaceuticals in wastewater from households, livestock farms, hospitals and pharmaceutical manufactures. *Chemosphere*, 82(2), 179–186. doi:10.1016/j.chemosphere.2010.10.026 PubMed

Sourav, M., & Alok, S. (2014). Treatment of Pharmaceutical Waste with Special Emphasis to Treatment Processes - A Review. *International Journal of Environmental Research and Development*, 4(2), 171–176.

Stackelberg, P. E., Gibs, J., Furlong, E. T., Meyer, M. T., Zaugg, S. D., & Lippincott, R. L. (2007). Efficiency of conventional drinking-water-treatment processes in removal of pharmaceuticals and other organic compounds. *The Science of the Total Environment*, 377(2-3), 255–272. doi:10.1016/j.scitotenv.2007.01.095 PubMed

Thompson, J. M., Gundog̃du, A., Stratton, H. M., & Katouli, M. (2013). Antibiotic resistant *Staphylococcus aureus* in hospital wastewaters and sewage treatment plants with special reference to methicillin-resistant *Staphylococcus aureus* (MRSA). *Journal of Applied Microbiology*, 114(1), 44–54. doi:10.1111/jam.12037 PubMed

Trovó, A. G., Melo, S. A. S., & Nogueira, R. F. P. (2008). Photodegradation of the pharmaceuticals amoxicillin, bezafibrate and paracetamol by the photo-Fenton process - application to sewage treatment plant effluent. *Journal of Photochemistry and Photobiology A Chemistry*, 198(2-3), 215–220. doi:10.1016/j.jphotochem.2008.03.011

Tumeo, E., Gbaguidi-Haore, H., Party, I., Bertand, X., Thouverez, M., & Talon, D. (2008). Are antibiotic-resistant *Pseudomonas aeruginosa* isolated from hospitalised patients recovered in the hospital effluents? *International Journal of Hygiene and Environmental Health*, 211(1-2), 200–204. doi:10.1016/j.ijheh.2007.02.010 PubMed

Vieno, N. M., Hrkki, H., Tuhkanen, T., & Kronberg, L. (2007). Occurrence of pharmaceuticals in river water and their elimination in a pilot-scale drinking water treatment plant. *Environmental Science & Technology*, 41(14), 5077–5084. doi:10.1021/es062720x PubMed

Viessman, W., & Hammer, M. (1993). *Water Supply and Pollution Control*. Harper Collins College Publishers.

Vlyssides, A. G., Loizidou, M., Karlis, P. K., Zorpas, A. A., & Papaioannou, D. (1999). Electrochemical oxidation of a textile dye wastewater using a Pt/Ti electrode. *Journal of Hazardous Materials*, 70(1-2), 41–52. doi:10.1016/S0304-3894(99)00130-2 PubMed

Watkinson, A. J., Micalizzi, G. B., Graham, G. M., Bates, J. B., & Costanzo, S. D. (2007). Antibiotic-Resistant *Escherichia coli* in Wastewaters, Surface Waters, and Oysters from an Urban Riverine System. *Applied and Environmental Microbiology*, 73(17), 5667–5670. doi:10.1128/AEM.00763-07 PubMed

Xiaosong, C., Michael, T., Meyer, X., Liu, Q., Zhao, H., Chen, J., Chen, Z., Qiu, L., Yang, J. C., & Weiqun, S. (2010). Determination of antibiotics in sewage from hospitals, nursery and slaughter house, wastewater treatment plant and source water in Chongqing region of Three Gorge Reservoir in China. *Environmental Pollution*, 158(5), 1444–1450. doi:10.1016/j.envpol.2009.12.034 PubMed

Yamina, B., Tahar, B., Lila, M., Hocine, H., & Laure, F. M. (2014). Study on Cadmium Resistant-Bacteria Isolated from Hospital Wastewaters. *Advances in Bioscience and Biotechnology*, 5(08), 718–726. doi:10.4236/abb.2014.58085

Zhang, G., Ji, S., & Xi, B. (2006). Feasibility study of treatment of amoxicillin wastewater with a combination of extraction, Fenton oxidation and reverse osmosis. *Desalination*, 196(1-3), 32–42. doi:10.1016/j.desal.2005.11.018

Zheng, T., Wang, J., Wang, Q., Meng, H., & Wang, L. (2015). Research trends in electrochemical technology for water and waste water treatment. *Applied Water Science*, 1–18.

Zhimin, Q., John, J. M., Melanie, R. M., Rao, S., & Craig, D. A. (2006). Treatment of Antibiotics and Antibiotic Resistant Bacteria in Swine Wastewater with Free Chlorine. *Journal of Agricultural and Food Chemistry*, 8144–8154. PubMed

ADDITIONAL READING

Liu, M., Zhang, Y., Ding, R., Gao, Y., & Yang, M. (2013). Response of Activated Sludge to the Treatment of Oxytetracycline Production Waste Stream. *Applied Microbiology and Biotechnology*, 97(19), 8805–8812. doi:10.1007/s00253-012-4589-8 PubMed

Miller, J. H., Novak, J. T., Knocke, W. R., Young, K., Hong, Y., Vikesland, P. J., Hull, M. S., & Pruden, A. (2013). Effect of Silver Nanoparticles and Antibiotics on Antibiotic Resistance Genes in Anaerobic Digestion. *Water Environment Research*, 85(5), 411–421. doi:10.2175/106143012X13373575831394 PubMed

Pasquini, L., Merlin, C., Hassenboehler, L., Munoz, J.-F., Pons, M.-N., & Görner, T. (2013). Impact of Certain Household Micropollutants on Bacterial Behavior. Toxicity Tests/Study of Extracellular Polymeric Substances in Sludge. *The Science of the Total Environment*, 463–464, 355–365. doi:10.1016/j.scitotenv.2013.06.018 PubMed

Roger, J. A. (2018). Advances in multisensor information fusion: A Markov–Kalman viscosity fuzzy statistical predictor for analysis of oxygen flow, diffusion, speed, temperature, and time metrics in CPAP. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, 32(4), 234–246. doi:10.1111/exsy.12270

Rongjun, S., Guangshan, Z., Peng, W., Shixiong, L., Ryan, M. R., & Crittenden, J. C. (2015). Treatment of Antibiotic Pharmaceutical Wastewater Using a Rotating Biological Contactor. *Journal of Chemistry*, 705275. Advance online publication. doi:10.1155/2015/705275

KEY TERMS AND DEFINITIONS

Chemical Oxygen Demand: Amount of oxygen required to oxidize both organic and inorganic substances in wastewaters.

Electrochemical Treatment Technique: Any process either caused or accompanied by the passage of an electric current and involving in most cases the transfer of electrons between two substances—one a solid and the other a liquid

Electrocoagulation: Is a broad-spectrum treatment technology that removes total suspended solids (TSS), heavy metals, emulsified oils, bacteria and other contaminants from water (is a technique used for wash water treatment, wastewater treatment, industrial processed water, and medical treatment).

Electrodeposition: Is a process that uses an electric current to reduce dissolved metal cations so that they form a thin coherent metal coating on an electrode.

Electroflotation: Is a process of floating of pollutants to water surface by tiny bubbles of hydrogen and oxygen gases generated from water electrolysis. Electroflotation module consists of electroflotator with insoluble electrodes, tanks for alkali and flocculant solutions, pumps, rectifier, sludge collecting system (Skimmer).

Electrophoretic Deposition: Is a term for a broad range of industrial processes which includes electrocoating, cathodic electrodeposition, anodic electrodeposition, and electrophoretic coating, or electrophoretic painting.

Emerging Pollutants (EPs): Synthetic or naturally occurring chemicals that are not commonly monitored in the environment, but which have the potential to enter the environment and cause known or suspected adverse ecological and (or) human health effects.

Ozonation Treatment: Oxidation process that requires the use of ozone.

Design and Implementation of a Location-Based Service With Emphasis on a Geographical Database

Wen-Chen Hu

University of North Dakota, USA

INTRODUCTION

Smartphones are the most popular electronic devices nowadays. People use them to perform daily activities, like managing messages and emails, browsing the mobile Internet, and watching video, anywhere and anytime. One of the popular apps is location-based service (LBS), which is a service based on the geographical position of a mobile handheld device. One of the LBS examples is to find a nearby ethnic restaurant by using a smartphone. According to the market research reports, (i) the global LBSs market was valued at \$11,994 million in 2016, and is expected to reach \$61,897 million by 2022, supported by a growth rate of 26.6% (Malani, 2017), and (ii) the growing trend for the integration of location-based search with social networking websites will post a growth rate of close to 40% for the global LBS market during 2017-2021 (Technavio, 2017). They show the high importance and popularity of LBSs. However, building an LBS is not a simple task. This research tries to help developers by showing how to build an LBS.

Building a location-based system (LBS) is not easy because various disciplines and subjects are involved. This article tries to mitigate the high complication and difficulty by showing how to build an LBS from the ground up with a focus on connecting the mobile devices to a server-side geographical database. The five major components of an LBS system and our methods for implementing them are listed below:

- *Mobile handheld devices:* Android emulator (AVD), instead of a smartphone, is used for development.
- *Positioning systems like satellites:* The DDMS (Dalvik Debug Monitor Server) of an AVD is employed to send the mock locations, instead of using actual road testing.
- *Mobile and wireless networks:* URL connection, provided by Android, is used to connect the emulator to the service provider.
- *Service providers:* The LBS is to show a direction between the user and the nearest participant.
- *Geographical databases:* A small geographical database is used by the service provider.

Readers, who are interested in LBS research and applications, can use this paper to build their LBS prototypes. Based on the prototypes, they can realize their LBS research and show the advantages of their proposed methods.

The rest of this paper is organized as follows. Section 2 introduces related software, tools, and services including (i) Android and Android Studio, (ii) MySQL and MySQL Workbench, and (iii) various location-based services. The proposed LBS app is introduced in Section 3, which includes three

DOI: 10.4018/978-1-7998-3479-3.ch094

sub-sections: (i) the proposed system, (ii) the server-side geographical database used by the proposed system, and (iii) the result screenshots. Section 4 explains how to build the proposed system including three sub-section: (i) Android server connection, (ii) client-side code, and (iii) server-side code. The final section summarizes this study and gives possible LBS projects.

BACKGROUND

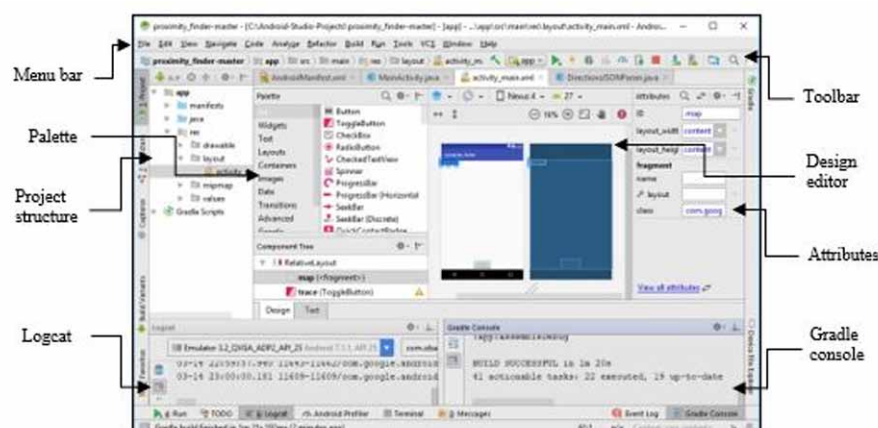
This section introduces the software and tools used to build the proposed system including (i) Android and Android Studio and (ii) MySQL databases and MySQL Workbench. In addition, various location-based services are listed to give readers an idea what the market trend of LBSs is.

Android and Android Studio

Android (Android, n.d.a) is the most popular mobile operating system as it is used by more than 80% of smartphones in 2017 (IDC, 2018). It is a software stack based on the Linux kernel for mobile devices that includes an operating system, middleware, emulators, and various key applications like phone, browser, and contacts. It was initially developed by Android, Inc., which was acquired by Google, Inc. in 2005. Since then, Android has been revised and enhanced significantly and constantly for the ever-changing newer features of mobile devices. In the past, Android app development in Eclipse with ADT (Android Development Tools) was highly recommended. Since 2014, Android Studio (Android, n.d.b) became the primary IDE (Integrated Development Environment) for native Android application development. It is freely available under the Apache License 2.0 based on JetBrains' IntelliJ IDEA software. Figure 1 shows a user interface of Android Studio and its major functions are listed as follows:

- *Attributes*, providing controls for the selected view's attributes,
- *Design editor*, displaying the layout in both the design and blueprint views,
- *Gradle console*, a screen for managing the automated build system for the Android source code,
- *Logcat*, a screen showing the logs generated from the running programs,
- *Menu bar*, providing the essential functions of Android Studio,

Figure 1. A user interface of Android Studio



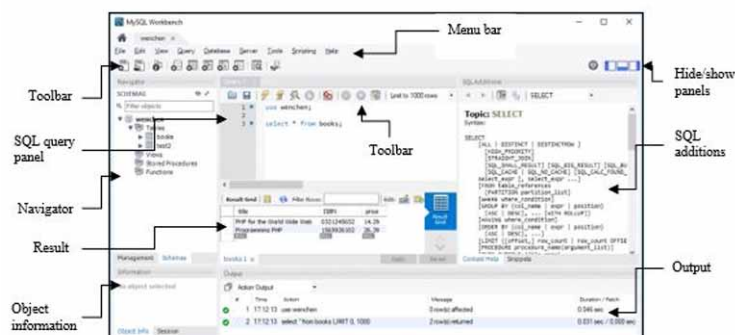
- *Palette*, providing a list of widgets and layouts that can be dragged into the layout in the editor,
- *Project structure*, showing the view hierarchy of the project, and
- *Toolbar*, providing buttons to configure the layout appearance in the editor and to change some layout attributes.

MySQL Databases and MySQL Workbench

The proposed geographical database uses MySQL (MySQL, n.d.a), an open-source relational database management system (RDBMS). It is supported by a company named MySQL AB, which is a division of Oracle Corporation now. MySQL is one of the major components of LAMP (Linux, Apache, MySQL, and PHP/Perl/Python) technology for website construction. It is reliable because tens of thousands of websites have used MySQL without incidences and, even better, it is free because it is open source. In addition, most programming languages like PHP and Perl are allowed to embed SQL (Structured Query Language) commands in their programs for adding, accessing, and processing data in MySQL. This research uses Android Java to connect to a server-side PHP program with embedded SQL commands to access the geographical database using MySQL. Other than using MySQL text user interface to manage MySQL, MySQL Workbench (MySQL, n.d.b) is a database IDE (Integrated Development Environment) for the MySQL database systems. It allows a DBA or developer to visually design, develop, manage, and maintain databases. Figure 2 shows a user interface of the MySQL Workbench and its major components are introduced as follows:

- *Hide/show panels*, changing the layout of the Workbench,
- *Menu bar*, providing the essential functions of MySQL Workbench,
- *Navigator*, showing the schema of the user,
- *Object information*, displaying the details of the object,
- *Output*, containing the log of executed commands,
- *Result*, showing output from the SQL commands in the query panel,
- *SQL additions*, including context help and SQL snippets for saving SQL code,
- *SQL query panel*, allowing users to enter SQL commands, and
- *Toolbar*, providing buttons to quickly manage the SQL commands in the query panel.

Figure 2. A user interface of MySQL Workbench



Location-Based Services (LBSs)

Various LBSs are available on the app stores and most major IT corporations provide their own LBSs. This sub-section introduces some of them, so readers can get a sense of what LBSs can affect our daily lives and implement their own LBSs based on this research:

- Foursquare (<https://foursquare.com/>), recommending interesting places like coffee houses, restaurants, and stores to users based on their previous browsing history, purchases, or check-in history.
- Uber (<https://www.uber.com/>), a form of sharing economy and of which major function is ride-sharing among its users.
- Google Maps (<https://maps.google.com/>), a web mapping service developed by Google. Other than the traditional map services like street maps and route planning for traveling, it also provides various recommendations such as restaurants and hotels.
- Bing Maps (<https://www.bing.com/maps/>), provided by Microsoft including the features of directions, traffic information, and people, business, and location search.
- iOS – Maps (<https://www.apple.com/ios/maps/>), a map or location-based service provided by Apple, Inc. including the features of indoor maps, navigation, and proactive suggestions.
- AutoNavi (<https://www.autonavi.com/>), a provider of digital map, navigation, and LBS in China offering products with features such as route planning, travel navigation, and offline maps from the Alibaba Group.

DESIGN OF THE PROPOSED LOCATION-BASED SERVICE (LBS)

Various LBSs can be found from the app stores. Examples of the services include finding nearby ethnic restaurants, navigation, and recommending interesting places or events. Other than the app user, most participants of the services are static or fixed; i.e., they, like the restaurants and concerts, are not mobile. This research investigates a kind of LBSs of which the participants are mobile. The potential of this kind of LBSs are great, but at the same time, their realization is different from the one of the traditional LBSs. This section introduces the design of the proposed system and implementation of the system will be given in the next section.

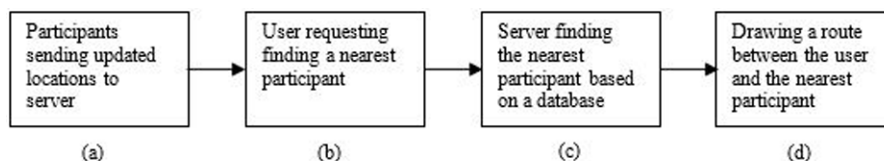
The Proposed System

The proposed location-based service is to draw a direction between the user and the nearest mobile participant. The major difference between the proposed system and other LBS systems is the participants of the former system are mobile. For example, a student is looking for his/her friends to borrow a calculator right before an exam. Traditional way to solve this problem is to find the nearest place where his/her friends live. However, the problem is his/her friends may be on the move and their locations may be different from time to time. The services would not be effective if the mobility feature is not considered. One way to keep track of the participants' locations is having the participants send their locations to the server once they move and the locations are saved in a server-side geographical database. The service then retrieves the current locations from the database and draws a direction between the user and the nearest participant. The following steps show how the proposed service works:

1. Participants sign up the service.
2. Entries are created in tables of a server-side geographical database.
3. Participants send their locations to the server if their locations change.
4. The database is updated according to the received location data.
5. A user submits his/her current location and a request to find a nearest participant.
6. Based on the requester's current location and the location data in the geographical database, the service draws a direction between the user and the nearest participant.

Figure 3 gives another way to explain the proposed service by giving a workflow of the proposed system. It includes the following four steps: (a) participants sending their updated location data to the sever, (b) the user looking for a nearest participant, (c) the server finding the nearest participant according to the locations of user and participants, and (d) drawing a route between the user and the nearest participant.

Figure 3. Workflow of the proposed system



The Proposed Server-Side Geographical Database

Geographical databases are databases storing geographic data including spatial data, like heights and locations, and attribute data like names such as McDonalds and functions such as restaurant. They are usually provided by a third party such as GeoNames (n.d.) because huge space is needed to store the data. This section proposes a small geographical database for facilitating geographic data storage, indexing, searching, and matching. This research uses a server-side MySQL (n.d.a) database to host the geographic data with an emphasis on mobile users. The database schema is shown in Table 1. It is made as simple as possible because the focus is on how to connect the mobile client to a server-side geographical database. The major difference between the proposed database and traditional databases

Table 1. Sample values of the proposed geographical database schema including (a) entities table and (b) LBS table

PID	Name	Function	Latitude	Longitude	Time	LBS	PID
1	Poke Mon	male	47° 55' 31" N	97° 2' 57" W	2019-02-09 21:21:40	1	1
2	Italian Moon	restaurant	47° 55' 40" N	97° 1' 60" W	2010-10-29 06:28:42	12	3
3	Super Mario	male	47° 55' 31" N	97° 1' 52" W	2017-05-28 10:51:30	2	4
4	GI Jane	female	47° 55' 60" N	97° 1' 98" W	2018-02-29 11:18:42	1	32
5	Valley Dairy	gas station	47° 55' 33" N	97° 2' 53" W	2017-03-37 06:21:43	4	120
6	Turtle River	park	47° 54' 72" N	97° 1' 70" W	1991-12-20 04:40:42	2	3
7	GF Bus	transportation	47° 55' 51" N	97° 1' 54" W	2019-02-54 10:33:21	1	3
8	Florida Sunshine	farmers' market	47° 55' 31" N	97° 1' 97" W	2018-06-29 19:21:44	...	...
(a)							(b)

is the entity could be mobile; i.e., its location could change from time to time. It consists of two tables: *entities* and *LBS* tables. The table *entities* include a column *time*, which is the time when the location is saved. The duration of the entity at the current location can therefore be found as the difference between the current time and the saved time. One of the functions of the duration is to check whether the location is valid. For example, if the duration of the location is more than a week, then the service may assume the mobile location is not valid. The other table *LBS* is used to find the participants of an instance of the service. One participant may join more than one instance of the service.

The Result Screenshots

This sub-section shows some of the screenshots from this service. It is to help readers understand what the service can achieve before they study the implementation details given in the next section. Instead of doing actual road tests, this research uses Android AVD (Android Virtual Device) to simulate the testing. Figure 4 shows three screenshots from this app. Figure 4.a shows the app in the Android Launcher, Figure 4.b is the control panel associated with the emulator, and Figure 4.c is the extended controls after clicking the item at the bottom of the control panel. The extended controls include a location function, which allows users to enter location data including latitude, longitude, and attitude and submit it to the app. Using this function is similar to the app receiving location data from the GPS (Global Positioning System).

Figure 4. (a) The proposed app shown in the Android launcher, (b) the control panel, and (c) the extended controls



Figure 5 shows an example of an application of this service, where the user is marked in Figure 5.a, Figure 5.b shows two more participants joining the service, and a direction between the user and the nearest participant is drawn in Figure 5.c. Another example is given in Figure 6, which is similar to the one in Figure 5, but with five participants.

Figure 5. (a) User marked, (b) two more participants joining the service, (c) direction between the user and the nearest participant



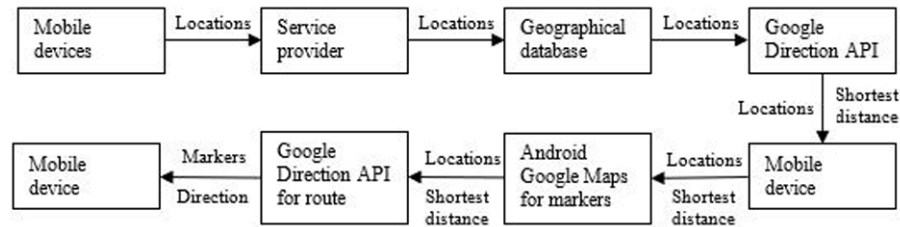
Figure 6. Another Example: (a) user marked, (b) five more participants joining the service, and (c) direction between the user and the nearest participant



IMPLEMENTATION OF THE PROPOSED SYSTEM

Design of the proposed system is given in the previous section and this section discusses the implementation part. Several programming languages are used in this implementation where Android Java is mainly used on the client-side for direction drawing, and PHP and Python are used on the server-side for accessing the geographical database and calling Google Maps Directions API (Application Programming Interface) (Google, n.d.) to find the shortest distance. Figure 7 shows a workflow of the implementation of the proposed system including (i) a user requesting a direction between him/her and the nearest participant, (ii) location data of the participants retrieved from the geographical database, (iii) the distances between the user and the participants calculated by using Google Maps Directions API, (iv) the data of the nearest participant sent to the user, and (v) the shortest direction drawn by using Google Directions API. Brief discussion of the implementation is given in this section. Further references for the related implementation can be found from the articles by the authors (Hu, 2018; Hu, Kaabouch, Yang, & Wang, 2014; Hu, Kaabouch, & Yang, 2014).

Figure 7. Workflow of the proposed system implementation



Android Server Connection

There are a couple of ways allowing Android devices to connect to the servers. Three of the classes for the connections are given as follows:

- `URLConnection`, which represents a communications link between the application and a URL,
- `HttpClient`, which is an interface for an HTTP client and was deprecated in android 5.1 and is removed from the Android SDK (Software Development Kit) in Android 6.0, and
- `HttpURLConnection`, which is the preferred method to connect to a remote server through HTTP.

This research connects a mobile client to the server via the `HttpURLConnection` class. The connection allows the client to send its location data to the server and receive the information in a JSON format the server offers. It is also used to connect to the Google Maps Directions server to request a route between the user and the nearest participant. Program 1 shows a code snippet used to connect to the server and receive data from the server.

```

private ArrayList<LatLng> downloadUrl(LatLng user_location) {
    HttpURLConnection urlConnection = null;
    URL url = new URL("ADDRESS_TO_SERVICE_SERVER_WITH_DEVICE_LOCATION");
    // Creating an HTTP connection to communicate with the URL
    urlConnection = (HttpURLConnection) url.openConnection();
    // Connecting to the URL
    urlConnection.connect();
    // Reading data from the URL
    while((line = br.readLine()) != null) {
        String[] latlong = line.split(",");
        double latitude = Double.parseDouble(latlong[0]);
        double longitude = Double.parseDouble(latlong[1]);
        LatLng loc = new LatLng(latitude, longitude);
        data.add(loc);
    }
}

```

Program 1. Code snippet for server connection

JSON stands for *JavaScript Object Notation*. It, much like XML, is syntax for storing and exchanging text information. This application parses JSON files returned from Google Maps Directions via web services and extracts map data from it. Program 2 shows a code snippet used to connect to the Google

Maps Directions service for drawing a direction between the user and the nearest participant. A URL for web services is built to request and retrieve the Google Maps Directions data in a JSON format. An example is given as follows:

```
https://maps.googleapis.com/maps/api/directions/json
?origin=37.422006,-122.084095
&destination=37.431098,-122.234912&sensor=false
private String downloadJSON(LatLng user_location, LatLng other_user_location)
{
    HttpURLConnection urlConnection = null;
    URL url = new URL("https://maps.googleapis.com/maps/api/directions/
json?origin=" +
        user_location.latitude      + "," + user_location.longitude + "&desti-
nation=" +
        other_user_location.latitude + "," + other_user_location.longitude +
        "&key=GOOGLE_API_KEY");
    // Creating an HTTP connection to communicate with the URL
    urlConnection = (HttpURLConnection) url.openConnection();
    // Connecting to the URL
    urlConnection.connect();
    // Reading data from the URL
    while ((line = br.readLine()) != null) { sb.append(line); }
```

Program 2. Code snippet of using web service for Google Maps Directions

The Client-Side Code

Program 3 shows a client-side code snippet for the following two major functions:

- Receive data from the server.
- Use the data to draw the direction between the user and the nearest participant and mark all participants.

The `doInBackground` method is used to request and retrieve the Google Maps Directions data in a JSON format by using web services. The `onPostExecute` method is used to parse the JSON string returned from the `doInBackground` method.

```
private class ChangeLoc extends AsyncTask <Location, Void, String> {
    @Override
    protected String doInBackground(Location... params) {
        mLatitude = params[0].getLatitude();
        mLongitude = params[0].getLongitude();
        final LatLng point = new LatLng(mLatitude, mLongitude);
        runOnUiThread(new Runnable() {
            @Override
```

```

    public void run() {
        mGoogleMap.moveCamera(CameraUpdateFactory.newLatLng(point));
        mGoogleMap.animateCamera(CameraUpdateFactory.zoomTo(12));
    }
});
ArrayList<LatLng> data = new ArrayList();
// Fetching the information of other users from web service
data = downloadUrl(point);
data.add(new LatLng(mLatitude, mLongitude));

// Drawing the marker, if destination location is not set
final ArrayList<LatLng> finalData = data;

runOnUiThread(new Runnable() {
    @Override
    public void run() { drawMarker(data); }
});
LatLng other_user_point = new LatLng(data.get(0).latitude, data.get(0).
longitude);
String results = "";
// Fetching the route from web service (Google Maps Directions)
results = downloadJSON(point, other_user_point);
return results;
}
// Executing in UI thread, after the execution of doInBackground()
@Override
protected void onPostExecute(String result) {
    super.onPostExecute(result);
    ParserTask parserTask = new ParserTask();
    // Invoking the thread for parsing the JSON data.
    parserTask.execute(result);
} // End of onPostExecute
}

```

Program 3. Client-side code snippet

The Server-Side Code

The server-side code includes the following two functions:

- Retrieve the participants' locations from the geographical database by using PHP.
- Find the shortest distance by using Python shown in Program 4.

Python and PHP are used because Python is convenient in text and array processing and PHP is recommended for web processing. The shortest distance is found by finding the distance between the user and each participant and comparing the distances.

```
users_locs = sys.argv[1].split(':'); # Receiving other users' locations
## Iterating for user locations
for distan in users_locs:
    url = 'https://maps.googleapis.com/maps/api/directions/json?origin='
        + origin + '&destination=' + distan + '&key=' + api_key
    ## Receiving the JSON object from Google Maps Directions
    f = urllib.urlopen(url)
    string = f.read()
    ## Parsing the JSON object
    j = J.loads(string)
    ## Receiving the distance
    dist = int(j["routes"][0]["legs"][0]["distance"]["value"])
    Distances.append(dist)
    ## Returning the user locations in an ascending order
for i in sorted(Distances):
    print users_locs[ Distances.index(i) ]
```

Program 4. Server-side code snippet

FUTURE RESEARCH DIRECTIONS

This research will have broader and important impacts on location-based research and services, as well as computer science education. Curriculum development and research initiatives will benefit from this research. This project involves a variety of topics and could be used to enhance curriculum development and for class projects. Readers may apply the knowledge and technologies learned from this article to the following suggested research or applications:

- Finding the interesting nearby places such as ethnic restaurants and movie theaters,
- Providing various recommendations such as concerts and museums including driving/walking routes,
- Using social networks to connect people on the move within a short distance,
- Location-based promotions and coupons such as sales and discounts, and
- Indoor positioning and navigation used to help users have a better visiting experience.

Other than the above suggested research areas, another critical issue about location-based services is security and privacy, which are considered difficult. For example, the service providers need the location data from the users in order to provide services. However, once the location data is shared with the providers, the user privacy may be violated. Without rigorous security and privacy, users would be reluctant to adopt the services. Related security and privacy research can be found in the Additional Reading section below.

CONCLUSION

10

Location-based services (LBSs), such as location-based social networks, navigation, and human behavior study and recognition, have already made a great impact on our daily lives. One of the major components of an LBS is the geographical database, which stores a huge amount of geographical data like restaurant and landmark information. The service providers use the data from the geographical databases to answer users' requests. This article pragmatically discusses the design and implementation of a location-based service, which includes five major components: (i) mobile handheld devices, (ii) positioning systems like satellites, (iii) mobile and wireless networks, (iv) service providers, and (v) geographical databases. The discussion covers the five components, but with emphasis on the geographical database. Readers will not have problems creating a location-based service after studying this article. Other than server-side databases, readers may consider using client-side databases like SQLite (n.d.) for mobile devices. However, the major problem of employing client-side databases is their storage sizes are usually not enough for geographical data.

REFERENCES

- Android. (n.d.a). *Android*. Retrieved February 18, 2019, from <https://www.android.com/>
- Android. (n.d.b). *Android Studio: The official IDE for Android*. Retrieved March 03, 2019, from <https://developer.android.com/studio/index.html>
- GeoNames. (n.d.). *GeoNames*. Retrieved March 10, 2019, from <http://www.geonames.org/>
- Google. (n.d.). *Google Maps Directions API*. Retrieved February 16, 2019, from <https://developers.google.com/maps/documentation/directions/>
- Hu, W.-C. (2018). *Data Engineering and Management*. Retrieved January 24, 2019, from <http://udncemcs01.und.edu/~wen.chen.hu/course/515/>
- Hu, W.-C., Kaabouch, N., & Yang, H.-J. (2014). Fundamental technologies and methodologies for mobile/handheld app development. In *Encyclopedia of Information Science and Technology*. IGI Global.
- Hu, W.-C., Kaabouch, N., Yang, H.-J., & Wang, X. (2014). Essential Android technologies and Google Maps APIs for location-based services. *Proceedings of the Midwest Instruction and Computing Symposium (MICS 2014)*.
- IDC. (2018). *Smartphone OS*. Retrieved March 02, 2019, from <https://www.idc.com/promo/smartphone-market-share/os>
- Malani, G. (2017, January). *Location-Based Services Market*. Allied Market Research.
- MySQL. (n.d.a). *MySQL*. Retrieved March 05, 2019, from <https://www.mysql.com/>
- MySQL. (n.d.b). *MySQL Workbench*. Retrieved February 16, 2019, from <https://www.mysql.com/products/workbench/>
- SQLite. (n.d.). *SQLite*. Retrieved August 13, 2018, from <https://sqlite.org/index.html>
- Technavio. (2017, May). *Global Location-Based Services (LBS) Market 2017-2021*. Technavio.

ADDITIONAL READING

Albelaihy, A., & Thayananthan, V. (2019, February). BL0K: A new stage of privacy-preserving scope for location-based services. *Sensors (Basel)*, 19(3), 696. doi:10.339019030696 PMID:30744047

Gartner, G. (Ed.). (n.d.). *Journal of Location Based Services*. Taylor & Francis.

Han, M., Li, L., Xie, Y., Wang, J., Duan, Z., Li, J., & Yan, M. (2018, March 28). Cognitive approach for location privacy protection. *IEEE Access: Practical Innovations, Open Solutions*, 6, 13466–13477. doi:10.1109/ACCESS.2018.2805464

Liu, B., Zhou, W., Zhu, T., Gao, L., & Xiang, Y. (2018, April 23). Location privacy and its applications: A systematic study. *IEEE Access: Practical Innovations, Open Solutions*, 6, 17606–17624. doi:10.1109/ACCESS.2018.2822260

Primault, V., Boutet, A., Mokhtar, S. B., & Brincoe, L. (2018). The long road to computational location privacy. *IEEE Communications Surveys and Tutorials*.

Wang, Y., Xu, D., & Li, F. (2016, June). Providing location-aware location privacy protection for mobile location-based services. *Tsinghua Science and Technology*, 21(3), 243–259. doi:10.1109/TST.2016.7488736

KEY TERMS AND DEFINITIONS

Android: Android, developed by the Open Handset Alliance led by Google, is a software stack for mobile devices that includes an operating system, middleware, and key applications.

Android Studio: Since 2014, Android Studio became the primary IDE (Integrated Development Environment) for native Android application development. It is free software under the Apache License 2.0 developed by Google. It includes the following features: (i) intelligent code editor, (ii) code templates and GitHub integration, (iii) multi-screen app development, (iv) virtual devices for all shapes and sizes, and (v) Android builds evolved, with Gradle.

Client-Server Architecture: It is a computing model including servers (the resource or service providers) and clients (the service requesters), where the server hosts, manages, and delivers the resources and services to be used by the client. For example, the electronic commerce systems are usually implemented by using a three-tier client-server architecture, where the server is the e-commerce provider (computing facilities) and the client is the end users (browsers).

Geographical Databases: They are databases storing a huge amount of geographical data such as information about restaurants and hotels. Service providers usually do not store and maintain all the information requested by users. Instead, geographical data and location information data are usually supplied by a third-party provider like a maintaining authority (e.g., mapping agencies) or business and industry partners (e.g., yellow pages).

Location-Based Services (LBSs): A location-based service is a service based on the geographical position of a mobile handheld device. One of the LBS examples is to find a nearby ethnic restaurant by using a smartphone.

MySQL Databases: They are open source relational database management systems (RDBMS) that use Structured Query Language (SQL), the most popular language for adding, accessing, and processing data in a database.

MySQL Workbench: MySQL Workbench, a unified visual tool, allows database architects, developers, and DBAs to perform various database management functions including data modeling, SQL development, user administration, and backup.

SQLite Databases: SQLite is an open source embeddable relational database management system. It supports most of the SQL-92 standard, has support for the major languages, and runs on all major operating systems. The SQL statements are compiled into assembly, which is executed in the SQLite virtual machine, the Virtual Database Engine (VDBE).

Climate Change: Inclusion of Gender and Cultural Diversity in Climate Change Actions

Mgbeodichinma Eucharia Onuoha Oragwa

Technical University of Mining Bergakademie Freiberg, Germany

INTRODUCTION

Climate change has influenced horticulture fundamentally, especially diminishments in soil water accessibility, impacts of hoisted CO₂ movements of warm and dampness limits to cropping, effect on drought, warm anxiety and different extremes, impacts on nuisances, weeds and ailments, and effects on soil fruitfulness. The question is no longer how much global warming and climate change rather what are the impacts and what ought to be done about it? To the later, the only way is by choosing to act with a sense of urgency starting from encouraging awareness of importance of gender and cultural diversity inclusion in climate actions, to energy policy reform down to mitigation and Adaptation measures (OECD/IEA, 2001).

However, Martin-Luther king's words four decade back still hold an intense reverberation in this twenty-first century where we are still faced with the savage urgency of a crisis that connects today and tomorrow; that is climate change crisis (UNDP, 2007/2008). *Climate change* is the characterizing human advancement issue of our age. All development is eventually about growing human potential and extending human freedom. It is about individuals building up the abilities that engage them to settle on decisions and to lead experience that they esteem. Climate change debilitates to dissolve human freedoms and limit choice. It raises doubt about the principle rule that human advance will make the future look better than the past (UNDP, 2007/2008). Moreover, since the peak of public mindfulness in the mid-1990s the global environmental issue known as climate change has been developing and changing quickly as of late (Sanderson et al., 2007). One of the most important factors, has been the inclusion of gender and cultural diversity in climate actions which in the past was either disregarded or given as an after-thought in the overall environmental change debate, while now it has become a more important hypothetical and policy issues (UNDP, 2011).

As climate is changing, culture is changing also (Wisner B., 2010). Across cultures, the effects of environmental change influenced women and men differently. Although women are excessively affected by environmental change, they assume a key role in adapting and mitigating climate change (UNDP Bangkok Regional Hub, 2019). For example, in the developing nations, women typically have the responsibility of gathering and generating food, collecting water and sourcing fuel for heating and cooking but with environmental change extrem events, these duties become harder to some women in the globe. Due to the above mentioned activities of some of these women in the globe, they are exposed with the knowledge and understanding of what is required to adapt to altering environmental circumstances and discovering practical solutions (UNDP Bangkok Regional Hub, 2019). *Culture* according to van den pol is generally defined as a total of meanings and knowledge that human beings required to operate in a

certain situation, for example, understanding of language, practices, rituals, views, values & norms etc. (Van den Pol, 2010).

Cultural diversity and climate change are two inseparable situations in the sense that, due to global warming: horticulture decreases and the generation of environmental refugees increases as a result of search for food and greener pastures, disruption of cultural continuity due to enhanced mortality from infectious diseases, mental dysfunction, stunted child development from malnutrition and the disruption of family life (Hoff, 2002). Knowing the importance of gender and cultural diversity, this article tends to promote more awareness to the importance of its inclusion on climate actions. Also note, that environmental change and climate change will be used in this paper as a synonym; the word 'Inclusion' also refers to 'consideration' in this paper.

BACKGROUND

As already mentioned above, in the developing nations, women typically have responsibility of gathering and generating food, collecting water and sourcing fuel for heating and cooking. With climate change, these tasks have become more harder due to extreme weather events such as diminishments in soil water accessibility, impacts of hoisted CO₂ movements of warm and dampness limits to cropping, effect on drought, warm anxiety and different extremes, impacts on nuisances, weeds and ailments, and effects on soil fruitfulness (UNDP Bangkok Regional Hub, 2019). To further discuss the connection between climate change, gender and climate finance, this section is divided into the following subsection:

Connectivity of Climate Change and Cultural Diversity

As already mentioned, to provide an appropriate solution to climate change, a cross-cultural communication is necessary as noted by Bernadet van den pol, since people from various countries, with various background, have to exchange their ideas and opinions about how to solve this problem (Van den Pol, 2010). As van de Pol is defining culture as a meaning or knowledge that human beings need to function in a certain situation, Keesing (1974) in his article is distinguishing between culture as an ideational system and culture as an integrated adaptive socio cultural system. Keesing further argued that as an ideational framework, culture is basically in the head of an aggregate mind and it only alludes to recognitions, convictions, standards and values. Culture in this setting can be utilized to explain social practices and their interrelation among thoughts and practices (Angula et al., 2014). For example, Angula et al. in their research observed that: the cultural beliefs, observation and view that perceive women as a frail and subordinate to a man still overwhelm the Namibian culture. Looking at the elements that upgrades vulnerability to environmental change such as: Decision making, access to information, assets, financial resources and technical skills, women in Namibia still have limited power at all degrees of administration (Angula et al., 2014).

However, climate change has various impact on culture. Climate change as known includes the interactions of numerous systems such as the air, hydrosphere, cryosphere and biosphere, each of which is enormously intricate in its possess right. The immediate desposition of rural people by huge scale projects attempted by government as a reaction to upgrade climate circumstances or clashes over water and arable land could be one cause of cultural interruption (Wisner B., 2010). Nevertheless, the combined consequences on water supply, illness patterns, cultivating framework and the habitability of seashores settlements have a cultural effects on the effected. Additionally, climate change has both irregular,

outrageous effects like enormous storms and heat waves, unavoidable, aggregate impacts like the annihilation of living things and the ascent of sea level. Both kinds of impacts as noted by Wisner have a role in dislodging human populace and interrupting their livelihoods and culture (Wisner B., 2010).

Also, irresistible disease pathogens like creepy crawlies and bacteria thrive in warmer atmosphere. The most affected are the vulnerable populations such as natives cultures, economically depleted groups and societies, and island countries. The variety of societies and human advancement that have developed over hundreds of years cannot be clarified without considering the impact of provincial ecological assets and overall local condition (Hoff, 2002). For instance: the atmosphere and soil, the type of water supplies, the biotic inhabitants of a local area, and other natural highlights that thoroughly formed the propensities, innovation, work, entertainment, travel, nourishment, and other cultural adaptations in uniquely various ways around the world are effected. Further instance where global warming impacts human culture is also found in the change of horticultural practices. For instance, unlike before, Asian civilizations were culturally based on rice; meso-American on Maize; the Peru on Potatoes, and European on wheat and other cereals; today the reverse is the case in the sense that production /manufacturing of food can no longer be describe as a culture to some extent (Hoff, 2002). These diverse effect on culture impacts cultural transmission particularly in rustic territories and among minority language bunches officially under stress due to economic globalisation, the urban float of the youthful and long period of disregard or even antagonistic vibe by those responsible of government language, training and cultural arrangements (Williams et al., 2013).

Connectivity of Climate Change and Gender

According to World Health Organization, *Gender* in the context of climate change is described as socially constructed norms, roles and relations that a given society considers appropriate for men and women. Gender determines what is permitted and valued in a woman or a man in a determined context (WHO, 2014). In the horticultural sector, provincial women in developing nations are the essential producers of staple nourishment, a sector that is profoundly bare to the dangers that accompany drought and unsure precipitation. In many nations, climate change implies that women and young girls need to walk further to gather water, particularly in the dry season. Women in Sub-Saharan Africa, as noted by UNDP go through 40 billion hours per year gathering water; and at the same time, expected to contribute much of the unpaid labor that will go into adapting to climate dangers through soil and water preservation, the building of anti-flood dikes, and increased off-farm business (UNDP, 2009).

Nevertheless, the negative effects of climate change can be felt in the present through catastrophic events such as avalanches, floods and hurricanes; and in the long haul, through increasingly progressive debasement of nature. The unfavorable effects of these events are already felt in numerous regions, including in regard to, entomb alia, farming and sustenance security; biodiversity and biological system; water assets; human wellbeing; human settlements and relocation patterns; and energy, transport and industry (UN Women Watch, 2009). In many of these cases as further described by UN Women Watch, women are more vulnerable to their effects than men due to the role they play in various regions of the globe. For instance: taking a gander at the territory of sustenance security, climate change has four noteworthy repercussions known, as described by UN Women Watch, namely: nourishment availability, nourishment accessibility, sustenance uses and nourishment frameworks dependability. Women agrarians presently account for 45-80% of all food manufacturing in the developing nations depending on the district; about 2/3 of these work powers are women in the developing nations and over 90% of them in

various African nations are occupied in agrarian work. Thereby making conventional sustenance sources to become increasingly capricious and scarcer (UN Women Watch, 2009).

The acknowledgement that vulnerability and reaction to environmental change impacts among men and women is not the same prompted the notion of gender dimensions in climate change dialogue and research. Effects of climate change are not gender unbiased; that is to say, gender roles are socially built and have generated disparity among men and women. The various roles that women and men play in societies and families are endangered to climate risks in numerous ways and this differentiates the vulnerability of women and men that are been affected by climate change. Additionally, the dependency of men and women on natural resources also make them vulnerable to climate change impacts (Angula et al., 2014).

Nonetheless, there is a significant interplay among environmental change, gender equality, women's empowerment, and accomplishment of the Millennium Development Goals. Environmental Change adaptation and mitigation endeavors legitimately influence advancement outcomes including issues of gender equality (UNDP, 2011). Moreover, policies that are educated by and respond to both women and men practical and strategic needs can lead to increasingly compelling climate responses as well as improvement across advancement goals. Beside, the design of the on-coming climate finance architecture plays a key role in whether or not, climate responses create sustainable development, including steps towards greater gender equality and women empowerment, or whether they inflame procedures that detriment women and strengthen historical or structural prejudice (UNDP, 2011).

Consequently, the World Health Organization records demonstrated that many of the health risks that are probably to be influenced by progressing climate change also show gender differentials. These distinctions are found in exposure to the aberrant and longer haul impacts of climate – related risks. For example, droughts in developing nations bring health risks through decrease accessibility of water for drinking, cooking and cleanliness, and through nourishment uncertainty. Women and their offspring unduly suffer health consequences of nutritional deficiencies and the burdens associated with travelling further to collect water (WHO, 2014). As a result of that, as noted by UNHCR, movement of populace given the ultimate changes and catastrophes could hinder and limit the conveniences of women and girls training as men emigration increment significantly living the burdens for their women and their children (Aguilar, et al., 2007).

Importance of Cultural Diversity consideration in Climate Change Actions

Having defined culture above, *diversity* is broadly defined to incorporate contracts in race, gender, national inception, ethnicity, capacity, and even geographical root. Cultural diversity are thus imparted through style, rules, shared significance, and even lingos, and the might be seen as contrasts by others in communication (Larkey, 1996). Native people are increasingly perceived to have impressive insight on issues identified with environmental change adaption. Studies have shown the value of indigenous peoples' perceptions of change in climate-related atmospher design, sea phenomena, and fire conduct. Their insight into past environmental motif can help remake historical benchmark. Cultural ecological insight of environmental health and species conveyances can contribute to customary suitable adaptation. Customary water-related insight, water collecting and cache have enabled indigenous people to endure in arid avalanche and adapt with drought for centuries (Williams et al., 2013). Van den Pol in his paper also says that the cultural differences matters alot since it influences both the content of the message and the way it is been expressed. Therefore, to provide an appropriate solution to climate change a cross-cultural communication is needed (Van den Pol, 2010).

O'Brien et al in their research denotes that values which is defined as interests, pleasure, likes, preferences, moral obligations, desires, wants, goals, needs, aversions, and attractions, play a significant role on how people respond to climate change action like mitigation and adaptation since it entails how people perceives the effects of climate change. In order words, how to adapt to climate change hinges on the values underlying people's perspectives on what the goals of adaptation should be (O'Brien et al., 2010). For instance, customary insight and practices as identified by Straza et al, are significant element of climate-related planning and of the reactions following climate occurrence and natural catastrophy and have a significant contribution to make since its incorporation in management planning can help to recognize how cultural components as well as religious practice and discord with versatility planning meets. In some other instances as seen in the case of Nambia above, women are kept from completely adding to climate-related planning, policy-making and execution. Although, women can be key advocate of ecological protection and climate versatility (Straza et al., 2018).

Furthermore, the use of customary insights by Research associations and western science in coping with and adapting to climate effects offer adavantages to both researchers and indigenous society (Williams et al., 2013). For instance, Analysts can benefit by procuring a better comprehension of indigenous qualities that should be obliged in adaption planning for landscape reclamation and management, customary advancement for adaptation, and local information, to supplement the technical learning base. While Indigenous societies can benefit by securing a better comprehension of environmental change, its effects and advancemnets for managing wit these effects that may lie outside of their cultures (Williams et al., 2013). Notwithstanding of all these benefits, concerns have been raised over potential unfriendly results of information exchange. Co-learning and co-generation procedures do not ensure fairness, equal standing or address control asymmetries since the natives are being welcomed to transfer customary information, often profoundly spiritual and core to third character, to take care of enormous -scale issues they cannot avoid and that are not of their creation (Williams et al., 2013).

Importance of Gender consideration in Climate Change Actions

Environmental change greatly effects those areas of the populace, in all nations, that are most dependent on natural assets for their occupation and who have minimal ability to react to natural dangers, such as dry seasons, avalanches, floods and cyclone. Women regularly face higher dangers and more difficulties from the effects of climate change in circumstances of neediness, although the majority of the world's poor are women as noted by UNFCCC (UN Climate Change, 2019). Knowing that women assume a demanding role in response to climate change due to their provincial expertise of and leadership in e.g. feasible resource administration and leading feasible practices at the family and society level, their involvement at the political level results in more responsiveness to citizen's needs, frequently expanding participation across party and ethnic lines and conveying more sustainable harmony (UN Climate Change, 2019). For instance, at the provincial level, women's involvement at the leadership level has prompted improved results of climate related projects and actions. On the contrary, if procedures or projects are executed without women's meaningful involvement, it can expand existing imbalances and diminishing viability (UN Climate Change, 2019).

This is one of the reasons why Parties to the UNFCCC have perceived the significance of including women and men equally in UNFCCC procedures and in the development and execution of national climate actions that are gender-responsive by setting up a devoted plan item under the Convention tending to issues of gender and environmental change and by incorporating all-encompassing content in the Paris Agreement (UN Climate Change, 2019). This Paris agreement produced a new fund called GCF that is to

replace the Kyoto protocol by 2020. Green Climate fund is a type of climate change finance established with the goal to mobilize climate fund to help scale-up mitigation and adaption action in developing nations on equal measure (Briefing Note on Paris Agreement V., 2016). Whereas climate finance is the exchange of funds from developed nations to the developing nations to empower them to adapt to the unavoidable effect of climate change by reducing GHG and embarking on clean energy advancement ways (Foe Briefing, 2010). The Green climate Fund is seen as an instrument that can advance gender correspondence by its structures and working strategies. In order words, an efficient gender delicate approach in the GCF will guarantee that both women and men profit from intercessions funded by the GCF through expansion of involvement, recognizing potential execution bottleneck, and upgrading the effort of the activities (Schalatek et al., 2013).

Moreover, according to United Nations climate change, women frequently demonstrate more interest for the environment, support pro-environmental actions and vote for pro-environmental leaders, their greater participation in politics and in nongovernmental organizations has resulted in environmental achievements, with multiplier impacts across all the Millennium Development Goals. Thereby calling for upholding women's involvement through legislation and policies that pledge that women will be heard and participate in decision-making (UN Climate Change, 2016). Evidences have also showed that women empowerment and advancing gender equality in the issue of climate change can render results across different sectors, including food and economic security and health. And that it will also lead to a more environmentally friendly decision making at household and national levels (IUCN, 2015). However, devoting more resources into inclusive, multilateral and multisector Climate Change Gender Action Plans can assist nations to advance comprehensive action that incorporate gender interests and expands on women's rare knowledge and viewpoints. And at the same time, ensuring equal space and resources for women and men to participate in climate change decision making and action at all levels in all nations (UNDP Bangkok Regional Hub, 2019).

Furthermore, Winnie Byanyima – Director, UNDP Gender Team – said, if climate finance is wisely administered, it can help to transform traditional gender roles and reduce inequalities. Also, failing to incorporate women and men's diverse needs and perspectives into climate change responses will further disadvantage women and decrease policies' efficiency and effectiveness (UNDP, 2011). Gate also notes that gender inclusion in climate change actions matters because upto 2016 the world majority poor are women and girls who still live in unacceptable poverty (Gates, 2016). Consequently, women and men differ in their jobs, practices and dispositions with respect to activities that could assist to relieve climate change. Research shows that, in many nations men devour more vitality than women, especially for private transport, while women are frequently incharge for majority of the family consumption choices, including in relation to sustenance, water and family vitality. Considering all these information in decision and policy making could assist accomplish a more focus, adequate, healthy and climately friendly actions (WHO, 2014).

Gender inclusion simply means as indicated by UN women, utilizing co-benefits among gender correspondence and climate activity which requires a change in perspective that puts gender interests and the voice and organization of women and girls, and men and boys, at the focal point of environmental management endeavors and ventures (UN Women, 2016). In order to accomplish such a change in perspective, it is undeniable that creating climate finance gender – responsive is fundamental to empower such a move. Likewise, to achieve an upgraded adequacy of environmental change mitigation and adaptation mediation, gender equality concerns should be mainstreamed into the whole project cycle beginning with the identification of priority mediation down to accomplishment of the environmental management targets of a given jurisdiction or substance. *Gender mainstreaming* here simply refers to

thinking differently, adjusting environmental and development mediations so that they will profit men and women evenly, and changing social, economic and institutional structures towards gender fairness and women's empowerment in environmental activity and resilience building (UN Women, 2016).

Further Research

Since the themes of the encyclopedia are about organizational knowledge and technologies, gender affects and shape both of these themes. However, there is a lack of deep and analytical discussion about how gender really shapes organizational knowledge. In order words, who are winners and losers and in what contexts? Which groups of women in what particular contexts pay higher price for being empowered in climate change management? Why is climate Finance important to the climate change gender nexus? Can be researched and analysis in the further papers.

CONCLUSION

The impacts of climate change on Islands and rural areas have a critical ramification on society, culture and gender. Social and cultural gatherings are excessively impacted by environmental Change. These bands of people assume a significant role in sharing data and ought to be consulted and incorporated in all endeavors to create climate activity. Climate finance should be available to both men and women and intended to create mutual gains, not infuriate designs of disparity. Thereby guaranteeing more positive advancement outcomes and effectively advancing gender parity and women's empowerment, as well as other related social and economic contemplations.

It is also imperative to recall that women are not only defenseless to climate impacts but also active actors of change in regard to both mitigation and adaptation. Women frequently have a solid group of awareness and ability that can be utilized in climate change mitigation, hazard minimization and adaptation procedures. Due to their responsibilities in families and societies as stewards of natural and family resources, they are in best position to add to occupation procedures adjusted to changing climate realities. Nevertheless, women frequently rely on a scope of crop varieties to oblige climate fluctuation, but perpetual temperature change will not only lessen crop assortments but also conventional medicine choices. Therefore, inclusion of culture and gender scrutiny expands the viability and proficiency of environmental change policy and program advancement and builds resilience.

REFERENCES

- Aguilar, L., Araujo, A., & Quesada-Aguilar, A. (2007). Gender and Climate Change. *IUCN - The World Conservation Union*. Retrieved 18th July 2019 from, http://cmsdata.iucn.org/downloads/gender_fact-sheet_climatechange.pdf
- Angula, M., & Menjomo, E. (2014). Gender, Culture and Climate Change in rural Namibia. *Journal for Studies in Humanities and Social Sciences*, 3(1 & 2), 225–238. Retrieved July 26th, 2019, from http://repository.unam.edu.na/bitstream/handle/11070/1289/Angula_gender_2014.pdf?sequence=1&isAllowed=y
- Briefing, F. (2010). *International Climate Finance - An Overview*. Retrieved 17th October 2017 from, https://friendsoftheearth.uk/sites/default/files/downloads/climate_finance_us.pdf

Briefing Note on Paris Agreement V. (2016). Green Climate Fund and the Paris Agreement. In *Climate Focus Client Briefing Note on the Paris Agreement V* (pp. 1-5). Paris, France: Climate Focus. Retrieved 30th October 2017 from, http://www.climatefocus.com/sites/default/files/GCF%20and%20Paris%20Brief%202016.new_.pdf

Gates, B. a. (2016). *If You could Have One Superpower, What would It Be*. Gatesnotes: Retrieve 11th July 2019 from, <https://www.gatesnotes.com/2016-Annual-Letter>

Hoff, M. D. (2002). Climate Change, Human Systems, and Policy - Vol. 1 Effect of Global Warming on Human Cultural Diversity. *Encyclopedia of life Support Systems (EOLSS)*, 1 - 8. Retrieved 26th July 2019 from, <https://www.eolss.net/Sample-Chapters/C12/E1-04-03-03.pdf>

IUCN. (2015). Gender and Climate Change - Strengthening Climate Action by Promoting Gender Equality. *International Union for Conservation of Nature (IUCN) - Issues Brief*. Retrieved 21st June 2019 from, https://www.iucn.org/sites/dev/files/import/downloads/gender_and_climate_change_issues_brief_cop21__04122015.pdf

Larkey, L. K. (1996). Toward a Theory of Communicative Interactions in Culturally Diverse Workgroups. *Academy of Management Review*, 21(2), 463–491. doi:10.5465/amr.1996.9605060219

O'Brien, K., & Wolf, J. (2010). A Values-Based Approach to vulnerability and Adaptation to Climate Change. *WIREs Climate Change*. <https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.30>

OECD/IEA. (2001). *International Emission Trading from Concept to Reality*. IEA Publication.

Sanderson, J., & Islam, S. (2007). *Climate Change and Economic Development - SEA Regional Modeling and Analysis*. Palgrave Macmillan Ltd.

Schalatek, L., & Burne, K. (2013). *Operationalizing a Gender-Sensitive Approach in the Green Climate Fund*. Energia: International Network on Gender and Sustainable Energy. Retrieved 1st September 2017 from, https://us.boell.org/sites/default/files/schalatek_burns_gcf_gender-sensitive-approach.pdf

Straza, T. U., Lui, S., & Burfitt, B. U. (2018). Effects of Climate Change on Society, Culture and Gender Relevant to the Pacific Islands. *Pacific Marine Climate Change Report Card. Scientific Review (Singapore)*, 201–210. Retrieved July 19th, 2019, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/714525/13_Society_Culture_and_Gender.pdf

UNDP. (2007/2008). Human Development Report: Fighting Climate Change - Human Solidarity in a Divided World. New York: Palgrave Macmillan Ltd.

UNDP. (2009). Resource Guide on Gender and Climate Change. *United Nations Development Programme*. Retrieve 11th July 2019 from, www.undp.org.mx

UNDP. (2011). Ensuring Gender Equity in Climate Change financing. *United Nations Development Programme*. Retrieved 26th July 2019 from, https://www.undp.org/content/dam/undp/library/gender/Gender%20and%20Environment/EngendCC_7.pdf

UNDP Bangkok Regional Hub. (2019). Gender and Poverty and Climate Change Finance. *Governance of Climate Change Finance*. Retrieved 11th July 2019: <https://www.climatefinance-developmenteffectiveness.org/topic/gender-and-poverty-and-climate-change-finance>

United Nations. Climate Change. (2016). *Why Climate Change Is not Gender Neutral*. Retrieved 11th July 2019 from United Nations Framework Convention on Climate Change: <https://unfccc.int/news/climate-action-needs-gender-action>

United Nations. Climate Change. (2019). *Introduction to Gender and Climate Change*. United Nations Framework Convention on Climate Change: <https://unfccc.int/gender>

Van den Pol, B. (2010). *The Connection Between Culture and Climate change*. Retrieved 8th July 2019 from (C) DocPlayer.net: <http://docplayer.net/20899297-The-connection-between-culture-and-climate-change-by-bernadet-van-den-pol-1-introduction.html>

WHO. (2014). Gender, Climate Change and Health. *World Health Organization*. Retrieved 16th July 2019 from, <https://www.who.int/globalchange/GenderClimateChangeHealthfinal.pdf>

Williams, T., & Hardison, P. (2013). Culture, Law, Risk and Governance: Contexts of Traditional Knowledge in Climate Change Adaptation. *Climate Change*, 120, 531 - 544.

Wisner, B. (2010). Climate Change and Cultural diversity. *Blackwell Publishing Ltd*. <https://www.ncbi.nlm.nih.gov/pubmed/21132942>

Women, U. N. (2016). Leveraging Co-Benefits between Gender Equality and Climate Action for Sustainable Development. *UN Women - mainstreaming Gender Considerations in Climate Change Projects*. Retrieved 18th July 2019 from, https://unfccc.int/files/gender_and_climate_change/application/pdf/leveraging_cobenefits.pdf

Women Watch, U. N. (2009). Women, Gender Equality and Climate Change. *The UN Internet Gateway on Gender Equality and Empowerment of Women*. Retrieved 11th July 2019 from, http://www.un.org/womenwatch/feature/climate_change/

ADDITIONAL READING

Heyd, T. (2010). Climate Change, Individual Responsibilities and Cultural Frameworks. *Human Ecology Review*, 17(2). Retrieved November 21st, 2019, from https://www.jstor.org/stable/24707655?seq=1#page_scan_tab_contents

KEY TERMS AND DEFINITIONS

Climate Change: Alludes to any change in climate over time, whether due to regular inconstancy or on account of human activity.

Climate Finance: Is the exchange of funds from developed nations to the developing nations to empower them to adapt to the unavoidable effect of climate change by reducing GHG and embarking on clean energy advancement ways.

Culture: According to van den pol is generally defined as a total of meanings and knowledge that human beings required to operate in a certain situation, for example, understanding of language, practices, rituals, views, values and norms, etc.

Diversity: Is broadly defined to incorporate contracts in race, gender, national inception, ethnicity, capacity, and even geographical root.

Gender: In the context of climate change is described as socially constructed norms, roles, and relations that a given society considers appropriate for men and women.

Gender Inclusion: Simply means as indicated by UN women, utilizing co-benefits among gender correspondence and climate activity which requires a change in perspective that puts gender interests and the voice and organization of women and girls, and men and boys, at the focal point of environmental management endeavors and ventures.

Gender Mainstreaming: Here simply refers to thinking differently, adjusting environmental and development mediations so that they will profit men and women evenly, and changing social, economic and institutional structures towards gender fairness and women's empowerment in environmental activity and resilience building.

Green Climate Fund: Is a type of climate change finance established with the goal to mobilize climate fund to help scale-up mitigation and adaption action in developing nations on equal measure.

Values: Which is defined as interests, pleasure, likes, preferences, moral obligations, desires, wants, goals, needs, aversions, and attractions, which play a significant role on how people respond to climate change action like mitigation and adaptation.

Gendering Information and Communication Technologies in Climate Change

Sam Wong

University College Roosevelt, The Netherlands & University of Liverpool, UK

INTRODUCTION

There has been a wide range of literature exploring the relationships between information and communication technologies (ICTs), climate change and gender. The European Union (2018) regards ICTs as a revolution for developing countries. Food and Agriculture Organization (2018) considers ICTs as ‘transformative technologies. He highlights three specific functions of ICTs in tackling climate change: monitoring and measuring the impact of climate change, raising awareness and facilitating dialogue, and fostering ‘networked governance’ (p4). UNDP (2019) also suggests that increasing access to information and education is essential for reducing the unequal vulnerability of women to climate change.

The rosy picture of using ICTs to tackle gender inequalities in climate change interventions has, however, been under strong attack. O’Donnell and Sweetman (2018) criticise ICTs for focusing too much on technical solutions, and too little on the social dimensions of the problems. Castillo et al. (2018) also warn that climate change can magnify gendered and other forms of inequalities when resources and access to goods and services become constrained in a changing climate. Rising inequalities, FAO (2018) suggests, would also intensify the impacts of climate change for all individuals and communities.

Gender can be defined as the social expectations and stereotypes of how men, women, boys and girls, should behave in society. Gender enables some groups of men, and women, to gain access to ICTs, whilst constraining others from doing so (Heeks and Ospina 2010). Variations in levels of control over ICTs, built on unequal power relationships, have tremendous impact on how poor people adapt to changing climate and respond to disasters (Wong, 2019). Understanding the dynamic interactions amongst gender, ICTs, and climate change is crucial to development policies and gender empowerment.

The objectives of this chapter are to offer a systematic review of the complex ‘ICT-climate change-gender’ nexus. Drawing on examples and case studies in developing countries, it aims to identify the gaps in literature and to explore the mechanisms by which gender influences, and is influenced by, the ICT-climate change interactions. It also intends to highlight the positive and negative gendered outcomes of mitigation, adaptation and disaster-planning policies across sectors, such as forestry, agriculture and water.

The structure of this book chapter is as follows: it will first discuss the importance of gender in understanding the interplay between climate change and ICTs. It will then highlight some proposals which attempt to increase women’s visibility in the climate change and ICT policies. After examining the pros and cons of the proposals, the chapter will conclude by offering gender-sensitive interventions that help promote digital inclusion and e-governance.

BACKGROUND

10

Rashid (2016) point out that ICTs are not a new phenomenon in developing countries. Many not-so-poor communities already have television, telephone and radio for communication and entertainment. That said, the latest ICTs, such as mobile phones, internet radios, emails, blogs and videos, are potentially interactive and participatory. With these, information can be shared faster and more efficiently between individuals and communities. They reach a wider range of audience and help develop wider networks. In light of this, Kalas and Finlay (2009) suggest these new ICT developments can improve governance by ‘empowering the poor and marginalised to raise their voice for political accountability and concrete action’ (p9).

Gender is defined as the socially constructed roles and socially learned behaviours and expectations of women and men in a particular society (World Bank 2014). Gendered relations involve difference, inequality and power, and that shapes access to, and control over, material and symbolic resources (Phan *et al.* 2019). Gendered relations are ‘contextually specific and often changing in response to altering circumstances’ (Moser 1993, 230).

Gender matters in ICT and climate change research and practice because it highlights the differential access to information. It questions fundamental questions about the nature of information - what is information, who produces information, and how information is used (Roztocki *et al.*, 2019). Onwutuebe (2019) stresses that the access to, and the use of, ICTs is gender-shaped. Gender influences how the benefits, costs and risks of ICTs are distributed between, and within, groups of women and men. For instance, in some Ugandan schools, girls are expected to be discreet and should not run like boys. Influenced by this specific gender norm, along with limited supplies of computers in school, boys race to gain access to computers and the girls are left out (UN 2002, 5). Gender also influences people to make responses to a changing climate. For example, IDS (2008) suggests that it is often men, not women, who decide what investments are made on their farms – such as investments relevant to climate adaptation – since laws and social practices enable men to have control over land, money, credit and tools.

To theorise the ‘ICT-climate change-gender’ nexus, Gurumurthy (2004), Wong and Sharp (2009) and Ospina and Heeks (2010) have offered three different conceptual frameworks. Gurumurthy compares and contrasts five different theories: Liberal, Marxist, Eco-feminist, Third World, and Culture. Wong and Sharp use the ‘agency-institution-structure’ framework to analyse how elderly women negotiate access to cutting-edge water innovations. Ospina and Heeks, in contrast, build on vulnerability and the livelihoods framework to assess the ICT-gender relationship.

This chapter draws on ideas from each of these frameworks. It puts three key climate change policies at the centre of the analysis; they are: mitigation, adaptation and disaster management. By examining the processes of implementing ICT interventions in, as well as their impact on, these three climate change policies, it analyses the role of gender in influencing men’s, and women’s, access to resources. Our analyses are based on four gender perspectives:

Process of gendered negotiation: The process explores what, and how, information is created and communicated, and how men and women understand the meanings, expectations and experiences of their gendered roles. Gendered relations are constantly negotiated and contested. They enable some groups of men and women to obtain access to information and to influence the outcomes of climate change, while denying others the facility to do so.

Gendered outcomes: Men and women experience the benefits, costs and impact of the ICT-climate change interplay differently because of their differential power positions and social capabilities. The

gendered outcomes can be positive, negative and also unpredictable. They are assessed if, and how, people's practical and strategic needs are met, and if gender inequalities are challenged or reproduced.

Feedback: The gendered outcomes of ICT-climate change interactions provide feedback to existing gender relations. They allow men and women to re-negotiate their gendered roles and expectations.

Indirect influence: Some ICT-climate change interventions make an indirect impact on gender relations. Changing income distribution between men and women, for example, can affect their decision-making power.

We are, however, aware of the limitations of focusing solely on gender in our analysis. The access to ICTs is shaped, not only by gender, but also many other factors, such as socio-economic status, income, age and geography (Selwyn, 2004). Including the class analysis helps avoid the risk of homogenizing the diverse and complex experiences of women, and men, in understanding the relationships between climate change and ICTs. We are particularly interested in the gender-class intersectionality for three reasons: firstly, gender and class are mutually enmeshed. In Pollert's words, 'there are no ungendered class relations, and conversely, there are no gendered relations without a class dimension at some level' (1996, 646). Secondly, Cleaver (2003) argues that feminist studies tend to produce unnecessary binary divisions between 'powerful men' and 'powerless women'. She suggests that, while not all men are powerful, not all women are powerless. This analysis helps explain why poor, working class men are subordinated to the control of middle-upper class women in certain circumstances. Thirdly, both gender and class affect resource allocation and utilization for mitigation and adaptation. The 'gender-class' interactions offers explanations for digital inequalities, as well as solutions to digital inclusion (Masozera *et al.* 2007, 301).

MIGITATION

The mitigation strategies focus on how ICTs help reduce CO₂ emissions. In this section, we will pay attention to two prominent mitigation policies: Clean Development Mechanism (CDM) and Reducing Emissions from Deforestation and Forest Degradation (REDD+). We will examine if, and how, gender is sidelined in the process.

CDM was set up to ensure technology transfers, from developed to developing countries, in order to limit or prevent the occurrence of climate change. CDM is the largest source of mitigation finance to developing countries. ICTs, in the form of hardware, software and information, covering a wide range of new technologies in household energy, agriculture and food processing, forest management, and water pumping in rural areas, are part of the CDM programmes (Haworth *et al.* 2018).

CDM, on the whole, has been criticised for being 'top-down' in nature. IDS (2008) suggests that existing mitigation policies are over-technical and over-scientific, and leave little room for gender-focused work (p14). The limited consultation with local communities fails to 'give women and men an opportunity to define and have power over their livelihoods' (UNDP 2010a, 2). CDM is also accused of taking an implicit assumption of gender neutrality. UNDP (2010b) suggests that 'little research has been done on the gendered impacts of the CDM, making it impossible to discern the precise degree to which current projects have impacted women and men differently' (p3). Wamukonya and Skutsch (2002) explain that technology transfer in CDM is usually directed towards men, which reproduces the gendered norms that 'men are considered the decision-makers and the eventual managers of technology' (p121). A lack of recognition of women's contribution is also reflected in the CDM governance. Of the 61 people who have served as Executive Board members, UNDP (2010b) shows that only 14 were women (p3).

Another measure to reduce CO₂ is through REDD+. Deforestation is the second most important human-induced source of greenhouse gases (GTZ 2007). Tropical deforestation annually releases 1.5 billion tonnes of CO₂ into the atmosphere, and this accounts for 17% of the total of greenhouse gas emissions (ITU and GeSI 2007). Illegal logging, overgrazing, farming encroachment, rising population pressure, and exploitation of forest products by local communities for subsistence purposes are all responsible for forest degradation and a rapid reduction of tropical forest cover.

ICTs, supporting forest management and avoiding forest degradation, play a significant role in REDD+. Geographical information systems (GIS) and remote sensing images are used to monitor the geographical changes of forest cover over time. ICTs also help examine the change in quality of forests to avoid degradation on the ground. They help acquire information, such as soil type, topography, wildlife, growth and yield, to measure the changing stock (Nsita 2010). Skutsch *et al.* (2009) also suggest that ICTs can be used to promote community participation in forest management. By providing community members with hand-held ICTs, they can help count species in their forests, thus allowing 'inexpensive and bottom up monitoring initiatives to take place' (IISD 2011, 28)

The perception of danger, associated with forest monitoring, has some gender implications for local ICT forest surveillance. Drawing on examples in Usangu, Tanzania, Cleaver (2001) suggested that people form village defence committees to protect their community resources. The youth are recruited as foot-soldiers and the elders act as advisors on tactics. While women and girls are excluded from the forest policing teams, because the tasks are seen as dangerous and physically-demanding, the full responsibilities fall on the shoulders of men and boys.

While some groups of men take up the role to protect their forests, others can collude with illegal loggers, and that undermines the prosecution efforts. Studies by Nsita (2010) show that corrupt politicians, mainly men, take bribes from forest encroachers. Driven not only by financial incentives, but also by winning votes, especially during elections, the political leaders give the illegal dealers cover, such as altering forest reserve boundaries. All these would reduce the effectiveness of the ICTs in arresting illegal loggers.

The role of women and girls in forest conservation is also mixed. They are seen as partners as well as pariahs. In their participatory research in Indonesia, Haynes *et al.* (2010) suggest that, for the sake of daily survival, girls 'choose which trees to cut when [they] go to collect the firewood either after school or during holidays' (p23). However, they also help plant trees if seeds are available. Women are encouraged to join local forest committees to get their voice heard. However, Castren and Pillai (2011) find out that women are often under-represented in the committees. Their tokenistic participation also fails to make real impact on the decision-making around forest management.

ADAPTATION

Adaptation strategies aim to increase adaptive capacity of vulnerable groups by strengthening their coping strategies to withstand, recover from, and adapt to climate change (Gumucio *et al.*, 2018). There are three main roles of ICTs in adaptation. Firstly, aerial photographs, satellite imagery and grid technology predict, identify and measure the extent of the problems of climate change (ITU 2009). Weather projections and vulnerability assessments predict climate change impact, such as rainfall patterns. Locally-specific information feeds into decision-support tools, and these inform policy-makers and local stakeholders to enable them to come up with different adaptation and production strategies. Secondly, ICTs reduce the costs of information exchange and raise awareness and facilitate dialogue. For example, IISD (2011) suggests

that the expensive running costs of GIS data systems would be reduced by networks and information-sharing (p33). Thirdly, ICTs promote community-based adaptation exchange and citizen involvement. For instance, UNDP (2011) shows that participatory sensing by mobile phones and scenario planning workshops promote community-based solutions.

There is a sense of optimism in the e-adaptation literature that ICTs could open up an opportunity for previously-marginalised groups, especially poor women, to re-negotiate their roles in their communities. IDS (2008), for example, suggests that the process of selecting heat-tolerant plants and drought-resistant animal breeds would help women farmers assert their authority in their farms (p11). Nelson *et al.* (2002) also indicates that building resilience would improve farm productivity. Increasing women's incomes would result in economic and other forms of empowerment.

Nevertheless, Resurreccion (2011) argues that many e-adaptation interventions focus too much on technical solutions to reduce the impact of climate change, and too little on 'a complex set of responses to existing climatic and non-climatic factors that contribute to people's vulnerability' (p2). Men and women may re-negotiate their roles in the process of adaptation, but Ospina and Heeks (2010) warn that the process gives 'voice to the interests of certain groups that may not be the most vulnerable' (p30). In particular, women feel more constrained from building adaptive capacities because they are deprived of 'land rights, ownership rights for the means of production, technology, finances, information and training in climate adaptation' (Bathge 2010, 5). Without control over productive resources and access to loans, credits and agricultural extension services, Olson *et al.* (2010) argue that women do not have the power to decide what changes are needed in their farms. All these constraints, IDS (2008) suggest, reduce their incentives 'to engage in environmentally sustainable farming practices and make long-term investments in land rehabilitation and soil quality' (p4).

Nelson *et al.* (2002) also warn that many adaptation practices build on women's unpaid labour. It assumes that 'rural women are predisposed to taking an environmental care-tending role' (MacGregor 2010, 8). Their e-adaptation involvement could increase their workload, without properly addressing their strategic needs. In order to juggle increasing demands, women may shift more domestic responsibilities to their daughters, and that could reproduce inter-generational inequalities and also gender inequalities in the next generation (Oxfam 2011b).

DISASTER MANAGEMENT

ICTs facilitate disaster management by offering efficient emergency communication and post-disaster coordination (Chanana-Nag and Aggarwal, 2018). Satellite-based systems capture detailed meteorological data and offer more accurate weather forecasts. On the ground level, wireless sensor networks can monitor localised weather changes and can create a database of disaster history to construct trend- and pattern-analysis (IISD 2011). Local hazard mapping and vulnerability assessments help identify individual houses and farmlands at risk and illustrate evacuation routes. Early warning systems, especially before tsunami attacks, also help reduce casualties. Furthermore, mobile phones, text messaging and real-time on-line collaboration are useful for immediate response (Hussain 2010). Rapid data gathering, during the emergency response, facilitates logistics and provides decision-making support. ICTs also allow more flexible response to post-disaster reconstruction efforts. Additionally, ICTs provide detailed maps to show the extent and the spread of the diseases and rationalise allocation of resources for disease control (Ospina and Heeks 2010). Community radios also help raise public awareness and promote public health campaigns.

The effectiveness of ICTs in disaster planning, however, depends on who has, and uses this, access to information. Radio provides a simple and effective tool to transmit information in emergency, but research by Gurumurthy (2004) shows that it is often men, rather than women, who control the access to radio. In Gurumurthy's words, 'if the household has one radio, it is most likely to be used by men. Women may not have the leisure to listen to the radio, nor may be allowed to join the men sitting outside the house listening to radio' (p5).

In their participatory gender study in Indonesia, Haynes *et al.* (2010) found out that both male and female adults have negative perceptions of girls' responses to disasters. While they praise the boys for being able to evacuate the elderly and move livestock from the affected areas, they think 'girls can't escape from disaster' because they only know how to scream (p24). As a result, boys are allocated more resources in evacuation, and girls are left more vulnerable to the effects of the disasters.

Using the big floods in Bangladesh in 1991 as an example, Rohr (2005, quoted in IDS, 2008) highlights gender biases in the early warning systems. Rohr explains that warning information was rarely communicated to women because:

'..... warning information was transmitted by men to men in public spaces, but rarely communicated to the rest of the family and, as many women are not allowed to leave the house without a male relative, they perished waiting for their relatives to return home and take them to a safe place' (p7).

Early warning systems also need to address the cultural constraints that delay the timely escape of the vulnerable. Swimming and tree climbing, for example, are useful skills to escape flash floods, but women are not taught the same skills as their male counterparts (Brody *et al.* 2008). Women's clothes in Muslim countries could reduce their mobility. Many of them were drowned because of this in the Bangladeshi flood in 1991, whereas the men could cast off their outer garments without censure.

Apart from gender, class also offers explanations for different capacities of women, and men, in mitigation, adaptation and disaster management. Low-income social groups are more vulnerable to climate change than their high-income counterparts because they tend to live in more hazardous places and have less protection. Equally, the high-income groups are more quickly to recover from losses more quickly in natural disasters since they are more capable of drawing on insurance, social safety nets and entitlement programmes (Masozera *et al.* 2007). In term of social participation, Leduc *et al.* (2010) discover that indigenous Nepalese women living in mountain have more power than their counterparts in lowland. They have greater independence and more control in making decisions. Their relatively active involvement in managing household and community resources would, therefore, have significant implications for adaptation strategy deployment. In the case study of White Volta water management in North Ghana, Wong (2016) also indicates that wives of chieftains in the trans-boundary water committee agree to build a 10 meter-wide buffer zone to make better preparation for river floods. This ambitious land policy means that poor, working class female, and male, farmers have to stop cultivation along the river banks and to move uphill to seek alternative livelihoods.

SOLUTIONS AND RECOMMENDATIONS

The previous sections have examined how gender affects the effectiveness of ICTs in mitigation, adaptation and disaster management. They have also highlighted the obstacles that constrain some groups of women, in particular, from exercising their agency in drawing on ICT resources to tackle climate change.

In addressing the gender biases, some suggestions have been mentioned in the literature in an attempt to raise the visibility and representation of vulnerable women and men. The proposals are: collecting sex-disaggregated data, stressing women-only interventions, providing women-friendly funding, and focusing on the agency of women. This section will evaluate these proposals and examine how some suggestions may not be able to achieve women's strategic needs, and worse still, may further reinforce their subordination.

Gender mainstreaming aims to integrate gender analysis into project design and public policy-making. UNDP (2008, 3) defines gender mainstreaming as:

'The process of assessing the implications for women and men of any planned action, including legislation, policies or programmes, in all areas and all levels. It is a strategy for making women's as well as men's concerns and experiences an integral dimension of the design, implementation, monitoring and evaluation of policies and programmes in all political, economic and societal spheres, so that women and men benefit equally and inequalities is not perpetrated'.

The intention of gender mainstreaming is to avoid the absence of a clear ICT strategy and policy which could affect gendered outcomes (World Bank, 2018) and to ensure that gender concerns are incorporated from the beginning and not as a corrective afterwards (Hafkin and Taggart 2001).

The main obstacle to achieving gender mainstreaming in developing countries, Etta (2005) claims, lies in a lack of a clear set of sex-aggregated data over the use of ICTs and the impact of climate change interventions. UN (2011) also suggests that a systematic gender analysis requires a good set of sex-aggregated data in order to establish gender-sensitive benchmarks and indicators to measure changes and improvement.

While data matter in social research, IDS (2008) argues that a gender- and class-sensitive approach requires more than a set of disaggregated data since it risks reinforcing fixed binary roles assigned to women and men. Instead, good quality data should reflect the 'contested, changed and reinforced' processes of gender and social inequalities (Resuureccion 2011, 7). Qualitative information, such as local power structures and control over key economic assets and social norms, is crucial to 'evaluate women's situation as compared to that of men in relation to specific environmental concerns' (Lambrou and Piana 2006, 12). In addition, Gurumurthy (2004) stresses that engendering ICTs is not simply about greater use of ICTs by women or men, but it is about 'transforming the ICT system' to meet their strategic needs (p2).

Gender disparities in ICT and climate change policies, Mittal (2016) suggests, are a result of a lack of participation of the powerless in the decision-making process. Governance strengthening is, therefore, intended to identify and overcome barriers to gender participation in decision-making. To enhance gender involvement, two key questions are raised in the literature: firstly, who participates, and secondly, how the institutions are designed.

A key question touches on whether women are targeted exclusively in the interventions or whether men should equally be considered. The women-only advocates suggest that women are generally poorer than men, and consequently, they are more vulnerable to climate change. Women compares 70% of the world's poor (MacGregor 2010). More women live in rural areas than men, and they rely on natural resources for survival, yet, their land rights are often denied (Oxfam 2011b). In addition, the literacy rate of women is lower than that of men. In Sub-Saharan Africa, for example, only 53.3% of women are literate whilst men have achieved 71.2% (UNDP 2011). To enhance female participation, gender experts and female scientists are targeted and invited to join in the preparation of climate change policies in order to 'improve dialogue between IT and gender policy makers' (Hafkin and Taggart 2001, 5). A

quota system, such as fixing the numbers of female ministries and a percentage of women activities per district, is proposed to foster women's leadership in key institutions and gender-responsive funding for action (Jolly *et al.* 2004). Another women-only group example includes women's community gardens in Malawi to ensure food security in times of flooding, and women-only self-help groups to produce organic staple crops to cope with climate change (MacGregor 2010).

Women-only interventions have, however, been criticised for homogenising women's experiences. Middle class women, for example, usually have higher access to ICTs than most poor men in rural areas (World Bank 2014). UN (2002) also suggests that single mothers, widows, disabled women, aged women and women in urban and rural areas have different identities, and their strategies to adapt to climate change can be very different. Furthermore, not all women in developing countries are low in internet use. South Africa and Chile are two examples where have achieved 50% connection. Cleaver (2003) warns that the exclusivity of women also risks considering the power relationships between men and women a 'zero-sum game'.

FUTURE RESEARCH DIRECTIONS

Men are 'gendered beings' too (MacGregor 2010). Poor men and boys are also vulnerable to climate change. Older men, for instance, are less tied into social networks than women, and that would put them at a disadvantage when they seek help from their communities (IDS 2008). Understanding masculinity also helps understand how men and boys facilitate, as well as hinder, women and girls in building resilience and using ICTs. For example, in the 'hole of the wall' computer experiment in India, Hafkin and Odame (2002) noted that: 'the aggressiveness of boys pushing away girls prevented the girls from using the computers' (p13). Additionally, the trend of digital education in developing countries benefit middle class families more than their working class counterparts because the latter is less likely to send their children to schools since they do not consider education as investment.

Despite the seeming incompatibility, the women-exclusive interventions and the proposal to include men are not mutually exclusive. In many cases, women feel more comfortable expressing their concerns in meetings, without the presence of their husbands or other men. Thus, when UNDP (2010a) conducted vulnerability assessments in Morocco, men and women were consulted separately, but the interests of men and women were equally taken into account.

Another key question concerning gender-sensitive governance is about how institutional arrangements are made. Developing a close partnership with local NGOs seems to build better governance. NGOs offer ICT training and access in their outreach strategies and climate change-related capacity building programmes (UNDP 2011). However, UN (2002) finds that many women's organisations and gender-focused groups perceive ICT policies 'an issue reserved for specialists' (p10). A genuine bottom-up approach working with local stakeholders to develop ICT capacity is rarely materialised. Although the World Bank (2017) suggests that improving governance provides an opportunity to combine modern communication systems with indigenous knowledge, Gillard *et al.* (2007) find that indigenous knowledge is often ignored or unnoticed. The 'bureaucratisation of gender guidelines' and the prescribed initiatives, they argue, do not pay adequate attention to cultural specificity of gender relationships (p6).

How to finance ICT and climate change projects has significant implications for affordability, accessibility and gender. The World Bank (2018) suggests that privatisation and liberalisation help lower user-prices of ICTs. Reforming the telecommunication sectors in developing countries and the licensing of new operators to introduce competition, the Bank argues, would ensure ICT access to the poor and

the underserved. To finance broadband infrastructure, innovative public-private partnerships are seen as catalytic vehicles to attract additional private sector investment.

To expand women's financing options, UNDP (2011) suggests improving women's access to micro-finance and other sources of credit. It also encourages the development of 'gender-sensitive climate change investment plans' which will facilitate access to financing for those most in need, particularly women and community groups (p13). In Ghana, for instance, a specific fund was set up to support initiatives that target women and promote gender equality in adaptation strategies.

UN (2002) shows scepticism of privatisation. It suggests that: 'ICT policy is frequently driven by the politics of the ICT sector, often impelled by commercial interests' (p11). Oxfam (2011a) argues that expanding financing options by micro-finance would risk imposing greater financial burden and stress on vulnerable women and men. Oxfam instead stresses the need to set up 'gender-sensitive funding guidelines' and 'equal gender representation in all decision-making bodies' (*ibid*, 3).

The agency approach focuses on 'individual capability to be and to act' (Bebbington 1999, 5) and considers access to information an act of empowerment. To challenge structural constraints, Hakfin and Taggart (2001) stress the importance of active participation amongst women. Women should make use of SMS messaging technology to develop collective networks.

While raising women's agency is crucial to challenging gender stereotypes, the understanding of the structural arrangements is equally important in promoting gender sensitivity in the ICT-climate change nexus. The denial of women's rights to land, for example, underlines the structural constraints that women face in mitigating and adapting to changing climate (Gurumurthy 2004). The social embeddedness approach, exploring the interplay between agency and structure, seems to provide a more nuanced framework in understanding the mechanisms of how gender shapes, and is shaped by, ICT and climate change interventions. Ospina and Heeks (2010) adopt a similar approach, highlighting the combination of personal preferences, social pressure and other decision-making mechanisms that influence adaptive capability and the outcomes of *achievable* and *actual* functionings (p9-10, original emphasis). Women, for example, feel constrained from using public cybercafé and rural information centres, not only because cultural norms restrict their mobility, but also because of the perception of safety and the unsuitable opening hours of the shops (Bathge 2010).

CONCLUSION

This book chapter has examined how gender shapes, and is shaped by, the ICT-climate change interactions. It has demonstrated the unequal gender and social practices that mean women and girls are more likely to be systematically disadvantaged in terms of assets, institutions and structures in the 'ICT-climate change-gender' nexus. Firstly, compared to men, women are less likely to access technology information. Deprived of land rights, they cannot make decisions about whether, and which, agriculture-related ICTs would be implemented. In addition, women struggle to make their voice heard in the mitigation policies because of their restricted access to finance.

Secondly, women are more institutionally-constrained than men. Cultural norms of behaviour shape what decisions women, and men, can take and what responsibilities they are required to take on. The exclusion of women in forest governance, for example, is related to the perceived danger to women of using ICTs to monitor illegal logging activities. Men bearing risk to protect their forests are given a legitimate voice to make decisions on behalf of women. In disaster management, rigid gender norms and cultural practices delay the evacuation of women and girls in cases of emergency. That said, on some occasions, the heroic culture amongst men and boys also put them at risk.

Thirdly, in term of social structures, women are excluded from decision-making in the policy design and resource allocations. The effectiveness of ICTs in disaster prevention is undermined by the masculine culture of information dissemination. Women are also less represented in formal decision-making bodies, such as CDM and REDD+.

In our analysis, we are aware that women, and men, are not homogeneous groups. The gendered processes and outcomes of the climate change-ICT nexus are further complicated by other social forces, such as class. According to Roehr (2007), class identities play a role in explaining people's differences in influence, attitudes and contribution to climate change (p3). Our analysis has also suggested that middle- and high-income groups have better access to ICTs and higher capabilities in managing climate change than their low-income counterparts.

To conclude, the 'ICT-climate change' interventions need to be both gender- and class-sensitive by considering the following issues:

- (1) Contextualising gender mainstreaming: gender mainstreaming is useful for integrating gender analysis into ICT policies. It acknowledges that men and women perceive and receive information differently, and that requires diverse approaches to adaptation. However, we also need to take class identities into account. Any attempt to re-position women and girls as 'eco-carers' is problematic because this fails to capture their protective, as well as their destructive, role in relation to natural resources. Without addressing the unequal power relations between mother and girls, e-adaptive practices can also help reproduce the inter-generational inequalities.
- (2) Governance strengthening: crafting new and reforming old, institutional arrangements is essential to improve gender inclusion. Women-only interventions are sometimes necessary to empower previously-excluded women to engage in ICT-related decisions. However, poor and powerless men should also have their say in climate change and ICT strategies.
- (3) Gender-sensitive funding mechanism: securing adequate funding to support ICT interventions is crucial to gender empowerment. Yet, targeting women by micro-credit projects risks putting an additional financial burden on them, and that needs serious re-consideration.
- (4) Agency-structure dynamics: women are active agents, but they are socially constrained from engaging in ICT-related decisions. Women's class identities, preferences, institutional arrangements and politics need to be taken into account in order to tackle digital exclusion.

We argue that these four suggestions are useful for development agencies, governments and NGOs to improve the gendered and classed outcomes from use of ICTs in response to climate change.

REFERENCES

- Bathge, S. (2010). *Climate change and gender: economic empowerment of women through climate mitigation and adaptation*. Working Paper. Governance and Democracy Division, GTZ Governance Cluster.
- Bebbington, A. (1999). Capitals and capabilities: A framework for analyzing peasant viability, rural livelihoods and poverty. *World Development*, 27(12), 2021–2044. doi:10.1016/S0305-750X(99)00104-7
- Brody, A., Demetriades, J., & Esplen, E. (2008). Gender and climate change: mapping the linkages. A scoping study on knowledge and gaps. Prepared for the UK Department for International Development.

Castillo, G., Cohernour, E., & Barbera, E. (2018). Factors influencing technology use by Mayan women in the digital age. *Gender, Technology and Development*, 22(3), 185–204. doi:10.1080/09718524.2018.1558862

Castren, T., & Pillai, M. (2011). Forest governance 2.0. A Primer on ICTs and Governance. PROFOR Program on Forests (PROFOR). Washington, DC: World Bank.

Chanana-Nag, N., & Aggarwal, P. (2018). Woman in agriculture and climate risks: hotspots for development. *Climate Change*.

Cleaver, F. (2001). Institutional bricolage, conflict and cooperation in Usangu, Tanzania. *IDS Bulletin*, 32(4), 26–36. doi:10.1111/j.1759-5436.2001.mp32004004.x

Cleaver, F. (Ed.). (2003). *Masculinities matter! Men, Gender and Development*. Zeds Book.

Etta, F. (2005). Engendering ICT Policies in Africa. WSIS Gender Caucus effort. Washington, DC: World Bank.

European Union. (2018). *The 2018 PREDICT Key Facts Report. An Analysis of ICT R&D in the EU and Beyond*. Author.

Food and Agriculture Organization. (2017). *Information and Communication Technology (ICT) in agriculture. A report to the G20 agricultural deputies*. Food and Agriculture Organization of the United Nations.

Food and Agriculture Organization. (2018). *Gender and ICTs mainstreaming gender in the use of information and communication technologies (ICTs) for agriculture and rural development*. United Nations.

Gillard, H., Howcroft, D., Mitev, N., Richardson, H., & Ferneley, E. (2007). Shaping the Global Economy: Gender, ICTs and Development Research. *Proceedings of the 9th International Conference on Social implications of computers in developing countries*.

GTZ. (2007). *Reducing Emissions from Deforestation in Developing Countries. The Way Forward*. Climate Protection Programme.

Gumucio, T., Hansen, J., Huyer, S., & van Huysen, T. (2018). *Gender Responsive Rural Climate Services: A Review of the Literature. A Learning Agenda for Climate Information Services in sub-Saharan Africa*. USAID.

Gurumurthy, A. (2004). *Gender and ICTs: Overview Report*. Institute of Development Studies. University of Sussex.

Hafkin, N., & Odame, H. (2002). *Gender, ICTs and Agriculture*. A situational analysis for the 5th Consultative Expert Meeting of CTA's ICT Observatory meeting on gender and agriculture in the information society.

Hafkin, N., & Taggart, N. (2001). *Gender, Information Technology and Developing Countries: an Analytic Study*. Academy for Education Development. For the Office of Women in Development Bureau for Global Programs. Field Support and Research United States Agency for International Development.

Haworth, B., Biggs, E., Duncan, J., Wales, N., Boruff, B., & Bruce, E. (2018). Geographic information and communication technologies for supporting smallholder agriculture and climate resilience. *Climate (Basel)*, 6(4), 1–20. doi:10.3390/cli6040097

Haynes, K., Lassa, J., & Towers, B. (2010). *Child Centred Disaster Risk Reduction and Climate Change and Climate Change Adaptation: Roles of Gender and Culture in Indonesia*. Children in a Changing Climate Working Paper II. Brighton: IDS.

Heeks, R., & Ospina, A. (2010). ICTs, Climate Change and Development: an Emergent Field within a Changing Climate. Workshop Summary. Centre for Development Informatics, University of Manchester.

Hussain, G. (2010). *ICT in Agriculture and Disaster Management*. Bangladesh Agricultural Research Council.

IDS. (2008). Gender and Climate Change: Mapping the Linkages - A scoping Study on Knowledge and Gaps. BRIDGE, Institute of Development Studies, Brighton.

IISD (International Institute for Sustainable Development). (2011). *Africa Transformation-Ready: the Strategic Application of Information and Communication Technologies to Climate Change Adaptation in Africa*. IISD.

ITU and GeSI. (2007). *Using ICTs to Tackle Climate Change*. Available at: www.itu.int/dms_pub/itu-t/oth/4B/01/T4B10000010001PDFE.pdf

ITU (International Telecommunication Union). (2009). *ITU Symposium on ICTs and Climate Change*. ITU background and report.

Jolly, S., Narayanaswamy, L., & Al-Zu'bi, R. (2004). Gender and ICTs. Supporting Resources Collection. BRIDGE, Institute of Development Studies.

Kalas, P., & Finlay, A. (Eds.). (2009). *Planting the Knowledge Seed – Adapting to Climate Change Using ICTs. Concepts, Current Knowledge and Innovative Examples*. Building Communication Opportunities Alliance.

Lambrou, Y., & Piana, G. (2006). *Gender: the Missing Component of the Response to Climate Change*. Food and Agriculture Organisation of the United Nations.

MacGregor, S. (2010). Gender and Climate Change: from Impacts to Discourses. School of Politics, International Relations and Philosophy, Research Centre for the Study of Politics, International Relations and Environment, Keele University.

Mittal, S. (2016). Role of mobile phone enabled climate information services in gender-inclusive agriculture. *Gender, Technology and Development*, 20(2), 200–217. doi:10.1177/0971852416639772

Moser, C. (1993). *Gender Planning and Development: Theory, Practice and Training*. Routledge.

Nelson, V., Meadows, K., Cannon, T., Morton, J., & Martin, A. (2002). Uncertain predictions, invisible impacts, and the need to mainstream gender in climate change adaptation. *Gender and Development*, 10(2), 51–59. doi:10.1080/13552070215911

Nsita, S. (2010). *Current and potential use of ICTs in the forest sector governance in Uganda*. World Bank.

O'Donnell, A., & Sweetman, C. (2018). Introduction: Gender, development and ICTs. *Gender and Development*, 26(2), 217–229. doi:10.1080/13552074.2018.1489952

Olson, J., Rubin, D., & Wangui, E. (2010). *Gender, Agriculture and Climate Change. A Regional Analysis for USAID/East Africa*. United States Agency for International Development.

Onwutuebe, C. (2019, January-March). Patriarchy and women vulnerability to adverse climate change in Nigeria. *SAGE Open*, 9(1), 1–7. doi:10.1177/2158244019825914

Ospina, A. V., & Heeks, R. (2010). Unveiling the Links between ICTs & Climate Change in Developing Countries: A Scoping Study. Centre for Development Informatics, Institute for development Policy and Planning (IDPM), University of Manchester.

Oxfam. (2011a). *Climate Change Beyond Coping. Women Smallholder Farmers in Tajikistan. Experiences of Climate Change and Adaptation*. Oxfam Research Report. London: Oxfam.

Oxfam. (2011b). *Climate Change and Women Farmers in Burkina Faso. Impact and Adaptation Policies and Practices*. Oxfam Research Report. London: Oxfam.

Phan, L., Jou, S., & Lin, J. (2019). Gender inequality and adaptive capacity: The role of social capital on the impacts of climate change in Vietnam. *Sustainability*, 11(5), 1257–1276. doi:10.3390/s11051257

Rashid, A. (2016). Digital inclusion and social inequality: Gender differences in ICT access and use in five developing countries. *Gender, Technology and Development*, 20(3), 306–332. doi:10.1177/0971852416660651

Resurreccion, B. (2011). *The Gender and Climate Debate: More of the Same or New Pathways of Thinking and Doing?* Asia Security Initiative Policy Series Working Paper no.10.

Roztock, N., Soja, P., & Weistroffer, H. (2019). The role of information and communication technologies in socioeconomic development: Towards a multi-dimensional framework. *Information Technology for Development*, 25(2), 171–183. doi:10.1080/02681102.2019.1596654

Skutsch, M., McCall, M., Karky, B., Zahabu, E., & Peters-Guarin, G. (2009). *Case Studies on Measuring and Assessing Forest Degradation. Community Measurements of Carbon Stock Changes for REDD*. Forest Resources Assessment Working Paper No. 156, Forestry Department, Food and Agriculture Organisation of the United Nations.

UNDP. (2008). *Gender Responsive E-Governance: Exploring the Transformative Potential*. United Nations Development Programme.

UNDP. (2010a). *Millennium Development Goals and Climate Change Adaptation. The Contribution of UNDP-GEF Adaptation Initiatives towards MDGs*. Issue No.2: Advancing MDG3. United Nations Development Programme.

UNDP. (2010b). *Clean Development Mechanism. Exploring the Gender Dimensions of Climate Finance Mechanisms*. United Nations Development Programme.

UNDP. (2011). *Africa Adaptation Programme Experiences Gender and Climate Change: Advancing Development through an Integrated Gender Perspective*. Discussion Paper Series, Vol.1, March.

United Nations. (2002). *Gender Issues in ICT Policy in Developing Countries: an Overview*. Division for the Advancement of Women. EGM/ICT/2002/EP.1.

United Nations. (2011). *Fact Sheet: Women, Gender Equality and Climate Change*. The UN Internal Gateway on Gender Equality and Empowerment of Women. Available at: www.un.org/womenwatch

United Nations Development Program. (2019). *UNDP Gender Equality Strategy 2018-2021*. UNDP.

Wamukonya, N., & Skutsch, M. (2002). Gender angle to the climate change negotiations. *Energy & Environment*, 13(1), 115–124. doi:10.1260/0958305021501119

Wong, S. (2016). Can Climate Finance Contribute to Gender Equity in Developing Countries. *Journal of International Development*, 28(3), 428–444. doi:10.1002/jid.3212

Wong, S. (2019). Decentralised, off-grid solar pump irrigation systems in developing countries – are they pro-poor, pro-environment and pro-women? In P. Castro, A. Azul, W. Leal, & F. Azeiteiro (Eds.), *Climate change-resilient Agriculture and Agroforestry: Ecosystem Services and Sustainability* (pp. 367–382). Springer. doi:10.1007/978-3-319-75004-0_21

Wong, S., & Sharp, L. (2009). Making Power Explicit in Sustainable Water Innovation: Re-linking Subjectivity, Institution and Structure through Environmental Citizenship. *Environmental Politics*, 18(1), 37–57. doi:10.1080/09644010802624785

World Bank. (2014). *Information and communication technologies, poverty and development. Learning from experience*. World Bank.

World Bank. (2017). *ICT for greater development impact. World Bank Group strategy for information and communication technology 2012-2015*. World Bank.

World Bank. (2018). *Information and communication for development. Data-driven development*. World Bank.

ADDITIONAL READING

Holmelin, N. (2019). Competing gender norms and social practice in Himalayan farm management. *World Development*, 122, 85–95. doi:10.1016/j.worlddev.2019.05.018

Muchacha, M., & Mushunje, M. (2019). The gender dynamics of climate change on rural women's agro-based livelihoods and food security in rural Zimbabwe: Implications for green social work. *Critical and Radical Social Work*, 7(1), 59–72. doi:10.1332/204986019X15491042559655

Namukombo, J. (2016). Information and communication technologies and gender in climate change and green economy: Situating women's opportunities and challenges in Zambian policies and strategies. *Bottom of Form Journal of Disaster Risk Studies*, 8(3), 243–249. doi:10.4102/jamba.v8i3.243 PMID:29955317

Niebel, T. (2018). ICT and economic growth – comparing developing, emerging and developed countries. *World Development*, 104, 197–211. doi:10.1016/j.worlddev.2017.11.024

Nsengiyumva, G., Kagabo, D., & Gumucio, T. (2018) *Exploring pathways for gender-responsive climate services in Rwanda*. Poster presented at the Gender Summit-14 Africa, Kigali, Rwanda.

Rengalakshmi, R., Manjula, M., & Devaraj, M. (2018). Making climate information gender sensitive: Lessons from Tamil Nadu. *Economic and Political Weekly*, LIII(17), 87–95.

Renken, J., & Heeks, R. (2018). Social network analysis and ICT4D research: Principles, relevance and practice. *The Electronic Journal on Information Systems in Developing Countries*, 84(4), e12036. doi:10.1002/isd.12036

Wong, S. (2017). Gendering Green Climate Finance: Potential and Limitations. In T. Krüßmann & A. Ziegerhofer (Eds.), *Promoting Gender Equality Abroad. An Assessment of EU Action in the External Dimension* (pp. 116–137). LIT-Verlag.

KEY TERMS AND DEFINITIONS

Adaptation: Adaptation underlines the importance to help poor communities adapt to changing climate.

Disaster Management: Disaster management touches on how climate-related disasters, such as floods and droughts, are prevented or alleviated.

Gender: Gender is defined as the socially constructed roles and socially learned behaviours and expectations of women and men in a particular society.

Gender Mainstreaming: The process of assessing the implications for women and men of any planned action, including legislation, policies or programmes, in all areas and all levels.

Information and Communication Technology: ICTs include television, telephone, radio, mobile phones, internet radios, emails, blogs, and videos for communication, entertainment, activism and control.

Mitigation: Mitigation focuses on the processes to reduce CO₂ emissions.

Poverty: It does not simply mean a lack of money, but it also focuses on empowerment, autonomy, and dignity.

World Bank: Named as the International Bank for Reconstruction and Development, the World Bank was set up in 1945 with the aim of promoting the post-Second World War reconstruction. It has since evolved and become strongly focused on worldwide poverty alleviation.

Section 11

Ethics and Responsibility

Science and the Morality of Weapons Research

John Forge

University of Sydney, Australia

INTRODUCTION

If it were not for *weapons research*, there would be no predator drones or smart bombs or improvised explosive devices or assault rifles. The insurgents in the Middle East and elsewhere would have no means to fight, and there would be no wars, large or small. Even more importantly, there would be no vast arsenals of thermonuclear weapons capable of ending much of the sentient life on the planet. The world would then most certainly be a safer place. But weapons research is not something new: the gunpowder weaponry of the early modern period was the product of research, as were the torsion catapults in Greece at the time of Philip and Alexander of Macedon. Whatever else is true about weapons research, it is clear that it introduces new (or improved) means of killing and destruction, and this is sufficient to define the activity. For instance Forge has given the following, generally accepted, definition: Weapons research is research carried out with the *intention* of designing new weapons, or improving the design of existing weapons, or designing or improving the means for carrying out activities associated with the use of weapons (Forge, 2013: 14., emphasis added). Weapons research would appear to be a very weighty matter, something that one might imagine philosophers, and others who think about such things, would have had a lot to say; surprisingly, not much at all has been written on the subject, though some explanation of this neglect will be given in this chapter.

The main issue for *ethics* and weapons research centres on the ethical or moral evaluation of the activity: Is it *ever* morally justified to design the means to kill, harm and destroy, and if so, under precisely what circumstances? Turning to science and its relation to weapons research, the question here is the *role* that science plays in weapons research. Perhaps weapons research is a wholly (applied) scientific endeavour or perhaps science is a part of weapons research? Bringing ethics back in, if weapons research is deemed morally wrong, then is it the case that *whatever* role science plays is also wrong? To talk about science playing a role here means in practice that it is scientists, people, who are undertaking actions that appear morally suspect. To answer these questions, four examples will be given which will clarify the roles that science can play in weapons research. For instance, if weapons research itself is understood as applied science, as it is by Arrigo for instance (Arrigo 2000: 303), then one might expect this to entail the application of theory to the design for new weapons, for true or radical innovation.

There are other ways in which science can inform weapons research, as will be seen presently. The first of the three examples show how science, both as theory and as method, can directly inform weapons research in the sense that the intention of the work in question was to produce new or improved weapons. The fourth example is different. It concerns *dual use* research, work that is not motivated by the desire to produce weapons but nevertheless provides a basis for doing so. This is an issue of current concern, so is worth discussing here. Before moving on to these examples, some general, and very brief, comments about ethics and the way it can apply to an intellectual activity such as science are in order.

DOI: 10.4018/978-1-7998-3479-3.ch097

This is appropriate because it cannot be assumed that the audience for the present topic is familiar with philosophy or ethics; but it is necessary to have a framework.¹

BACKGROUND

A straightforward way to describe ethics is to say that it is a study which deals with what persons ought and ought not to do. It is thus to do with the choices, actions and behaviour of mature competent people. Some of the things that people do do not affect others, other humans, other sentient beings, in any significant way and hence these do not come under the purview of ethics. Those actions that do affect others are, however, open to moral or ethical evaluation: are they right or are they wrong? To resolve that question, one needs to appeal to a moral system. All such systems forbid certain actions, namely those that inflict unjustified harm on others. This is surely intuitive and obvious: *no one* wants to be harmed. It is almost by definition that no sentient being wants to feel pain – assuming that the pain does not indicate that some medical treatment is working or some such – and to be in pain is one form of being harmed.

Some moral systems require people not only to refrain from harming others but also to provide some positive benefit for them. Jeremy Bentham and John Stuart Mill, the nineteenth century English philosophers, famously believed that one ought to strive to increase the amount of happiness in the world. However, morality is supposed to be *impartial* in the sense that it forbids discrimination in regard to moral action. Prohibitions on harming do not end with family or friends or community or country: *nobody* should be harmed, no one at all. Some critics of the style of morality advocated by Bentham and Mill have pointed out that it is impossible to increase the amount of happiness in the world impartially: no one can make everyone happy! Just how serious this objection is is a matter of ongoing debate. But it is only necessary here to note that this kind of moral system shares the prohibition on harming with the former kind: for the topic at hand, it is clear that the moral evaluation of weapons research, whatever else it might involve, will not be such as to see it as an activity which aims to increase the amount of happiness in the world.

When a person does something that contravenes the dictates of morality, they are guilty of moral wrongdoing. If one person harms another, then this is obviously a bad thing from the perspective of the victim, but what of the agent, is it also somehow bad for her? The short answer is that it is bad for the agent, in the sense that performing immoral acts attracts the disapprobation and distrust of others, although this is not an easy idea to explain when space is limited, as it is here (see Forge 2008: 92-93) And we also need to ask if all actions that harm attract such moral censure or only those that are intentional. All intentional acts are such that the agent foresees the outcome, assuming that matters do turn out the way they were planned, but not all foreseen outcomes are intended. A dentist who drills her patient's tooth foresees that it will hurt, and in the sense harm the patient, but she intends to save the tooth not cause harm. This raises the issue of the justification of actions that knowingly but unintentionally cause harm. Some such action can be justified, but others cannot, as is the case for dual use research.

Most philosophers do not believe that the rules of morality are absolute and cannot be broken in any circumstances. For example, most accept that a moral rule such as “Do not cause pain” has justified exceptions. Clearly, a dentist who inflicts pain on her patient to save his teeth has not done something morally wrong – provided that the patient understands and assents to the treatment. Also, it is generally agreed that it is permissible to cause pain in self-defence, if that is the only way to defend oneself. This leads to the view that *justifiable* exceptions to the overall moral prohibition against harming will be such as to show that the harm inflicted will prevent other harms. Just how this is worked out will vary from

case to case, and it is here that much of the hard work in ethical reflection and evaluation takes place. One might think as a basic principle that the harms prevented should be at least as much or many or as great as the harms caused if there is to be justification, and that therefore it must be necessary to be able to make some informed assessment of what these might be. This brings the discussion back to weapons research, an activity that aims to provide the *means* to harm. It is now necessary to look at some examples.

Solutions and Recommendations

Examples of Weapons Research

1. It is well-known that both kinds of nuclear weapon, fission and fusion weapons, were the *direct results* of the application of scientific theory to design. Without advances in nuclear physics in the 1930s and 1940s, the very idea of a nuclear weapon would not have been dreamed up. Rhodes gives an excellent account of the genesis of the idea of a fission weapon, from the speculations of Szilard from 1933 to his work with Fermi in the US, and the ‘memorandum’ written by the émigré scientists Frisch and Peierls in England which made predictions about crucial nuclear parameters. The designs of the two kinds of fission bombs made during the Second World War were determined by the properties of the fissile materials used, matters that were uncovered by painstaking research.² The idea for a thermonuclear weapon emerged from the atomic bomb project, and was also driven by science and by scientists. Indeed, the fusion reactions that power these weapons must be initiated by ‘fission triggers’, so not only did the research into fission or atomic bombs lead to thermonuclear weapons, the products of the former were integral elements of the latter.³

The research leading to nuclear weapons was exceptional in that scientists, including all of those mentioned above, took the initiative in agitating for programmes to investigate the possibility of using the recently discovered fissile materials for a weapon of terrible destruction, to the subsequent regret of some of them. The military is now well-aware of the power of science and has set up agencies to oversee the application of scientific theory to the design of new weapons systems. The best-known is the US’s Defense Advanced Research Projects Agency, or DARPA; notice that, in common with armed forces the world over, the word “defense” appears in its name. There are equivalents in Russia, China, the UK, France, Israel and other countries.

2. In the nineteenth century the weapons that caused the majority of fatalities and injuries in war, artillery and the infantryman’s longarm, were transformed by the scientific analysis of existing armaments. Smoothbore cannon and muskets were transformed into steel, breech-loading guns whose recoil was damped and rifles firing clips of cartridges using smokeless powder.⁴ These are instances of science *improving* existing designs. One example will be given as an illustration.

All smoothbore weapons which fire balls, musket and cannon balls, are inherently inaccurate because of uncontrollable forces which act on balls as they spin in the air after leaving a gun barrel. This is called the Magnus Effect or Magnus Force, after the German physicist Heinrich Magnus who described it in 1852. It was also investigated by the English physicist Benjamin Robbins, in his book *The New Principles of Gunnery*, published a century before Magnus’ work (Steele 1994: 359). The effect is this: unless a ball only spins about the axis in the direction it is travelling, what is called the roll axis in aeronautics, it will experience unequal pressure on either side in the direction of its motion, and hence be deflected in the

direction of the axis of spin. In other words, it will not travel straight, and if it actually hits (anywhere near) its target, then this is a matter of luck.

This analysis showed that if only the bullet could be made to spin about the roll axis, accuracy would be greatly increased. The first attempts to do so involved ‘rifling’ musket barrels, fitting them with raised ‘lands’ spiralling down the inside of the barrel which caused the ball to spin about the required axis. There were two problems: the rifled musket was very slow to load because the ball had to be hammered down the barrel, and the barrel became fouled after only a few shots because the powder residue collected around the lands. These problems were solved by three innovations. There were, in the first place, advances in metal working that allowed relatively efficient breech-loading rifles – leakage of gas from around the breech was eventually dealt with by gunsmiths such as Paul Mauser. Then there was progressive improvement in propellant until eventually smoke-less powder became available. Finally, bullets were designed that would not stick in the barrel on the lands. The idea was that the bullet would have two parts; a base that was made out of a relatively soft material that would deform in the lands and impart a twist to a hard core made narrower than the base so that it would not stick in the barrel. After two attempts in England in the 1820s, eventually a French innovation by Captain Claude-Étienne Minié of the French Chasseurs was successful in 1847.⁵ This led to many orders of magnitudes increase in lethality: in the American Civil War over half a million casualties, including two hundred thousand killed, were caused by ‘Minié rifles’.

3. The application of scientific theory, either to produce quite new weapon designs or to improve existing ones, is a relatively new phenomena, beginning with the study of -ballistics in the seventeenth century (see Hall 2009). However, *systematic* research into weapons design took place at least as far back as the fourth century BCE. Remarkable work was done in that century in relation to the design of the torsion catapult, which marked the beginning of effective siege artillery. A torsion catapult works because of the restoring force exerted by ‘springy’ material, like rope, which has been stretched. Wrapping rope around a stout piece of wood fixed to a frame, embedding a handle in the mass of rope and then pulling the handle back will create a torque: the handle will fly back to its resting position when released. It can be therefore seen how, in principle, such a device can be used to cast stones when placed in a basket attached to the handle.

It is not known how the Greek engineers discovered the torsion principle and applied it to designing catapults, but this must have been done *empirically* as the theory was not formulated until much later. However, the design of torsion catapults was codified in treatises which gave general instructions how to build them. These instructions were general because formulae, known as the hole formulae, gave the key parameter for casting stones of a given weight or arrows of a given length. The Greek engineers had discovered formulae that related the optimal size of the hole in the catapult frame in which the wood about which the rope or springs was to be wound to the weight of the stone or length of the arrow. The former is more complex and involves a cube root, which gives an equation that Greek mathematics at the time could not solve exactly. Perhaps still more remarkably, the treatises gave approximation methods for solving the equations in question.⁶

Dual Use

4. To introduce this example it is necessary to begin with some words about the idea of dual use. To begin, three different categories of dual use items can be distinguished: research, technologies and

artefacts. These are clearly different sorts of things. Research is an activity, while technology is a form of knowledge, knowledge of the techniques for the production of artefacts, whereas artefacts are objects. Certain instances of these kinds are classed as dual use because they can have both a 'good' and 'bad' use. Ammonium nitrate, for example, is normally used as a fertilizer, but it can also be used to make so-called fertilizer bombs.⁷ The dual use categories are related: research aims to give us technology, which in turn produces artefacts. It is the latter that normally has the immediate impact on the individual, be this good or bad. Forge's suggested definition of a dual use item is as follows:

An item (knowledge, technology, artefact) is dual use if there is a (sufficiently high) risk that it can be used to design or produce a weapon, or if there is a (sufficiently great) threat that it can be used as an improvised weapon, where in neither case is weapons development the intended purpose. (Forge 2010: 117)

Threats differ from risks here in that they signify the intention to do something harmful, rather than the possibility that something harmful might happen. Risk is coupled here with the design of (new) weapons but threats with improvised weapons. States threaten one another, as Iran and Syria are doing at present, though the threat here is usually conditional. However, in regard to dual use, one is more concerned at the moment with sub-state actors, such as terrorist organisations, using or adapting existing items for improvised weapons, as has been done with ammonium nitrate, than with states arming themselves with improvised weapons, as states do not usually need to improvise weaponry.

There have been a number of issues about dual use that have been discussed in the literature, but what is of interest here is whether undertaking dual use research is itself morally wrong, or whether it is morally permissible – deliberately taking advantage of dual use research and developing weapons is morally wrong. The example to be used here has to do with the possibility of using dual use research to develop nuclear weapons. Recall that naturally occurring uranium comes in two isotopes: chemically identical species of atom that have different physical properties owing to the presence of differing numbers of neutrons. In the case of uranium, the lesser abundant U-235 is much more fissile, much more apt to fission and release energy, than the more abundant U-238. A sample of uranium metal is said to be enriched if the amount of U-235 is greater than 0.7%, which is the proportion found in nature. The greater the enrichment, the greater the neutron flux, the greater the flow of neutrons in the metal. It is possible in theory to make a nuclear weapon with uranium enriched to a little less than 20% but the accepted practical limit is set at 20%, what is referred to as highly enriched uranium or HEU, when 30 kg of the metal would be needed (Glaser 2017:8). By contrast, only 2.3 kg of 95% enriched uranium is needed for the same purpose – uranium enriched to 90% or greater is known as military or weapons grade. Civilian power reactors, on the other hand, only require enrichment to about 4%. A civilian power reactor fuel rod is therefore not itself a dual use item because it cannot be used to make a bomb, unless subject to further enrichment. Weapons grade uranium is not suitable for power reactors because the extremely high neutron flux that makes bomb making easy but cooling very difficult, so it is not itself a dual use item either. However, there is another sort of reactor, the research reactor, which does require HEU, and hence a sample of HEU is dual a use artefact: it can be used for research or bombs.⁸ The question then is whether a research project aimed at producing uranium enriched above 20%, but below weapons grade, designed for a research reaction is morally permissible, when there exists the possibility of the technology being used to produce the fuel for nuclear weapons.

Analysis and Discussion

11

If weapons research is morally wrong, then it would appear that the actions of any scientist who contributes to weapons research are also morally wrong. But the matter is not quite so straightforward. A scientist who knowingly, willingly and intentionally engages in weapons research is as open to moral censure as any other weapons researcher, assuming that weapons research is a proscribed activity. It is not so clear, however, that a scientist who does not intend that her work be used to further the ends of weapons research, but sees that it may well in fact do so is also blameworthy. It is less clear still that a scientist who does not foresee this outcome, assuming it comes about, is to blame. It would not be necessary to address any of these issues if weapons research was not morally wrong, so this matter must now be discussed.

It was mentioned above that the topic has received much less attention than it deserves.⁹ Forge, who has addressed the topic at some length and argued that weapons research is morally wrong under all circumstances, believes that the question of the morality of weapons research has been assimilated to questions about the justness of war, about which a great deal has been written of late, and he believes this is a mistake (Forge 2013: ix)). The consensus about war is that war is just or morally permissible – no blame attaches – if it resists aggression. This is the ‘modern view’ which sees just war as more than self-defence, other defence being allowed if the other is the victim of aggression. The implicit corollary is that measures such as weapons research are justified if they aid in the prosecution of a just war or in the pursuit of self-defence or deterrence.¹⁰ This is the standard rationale or *standard justification* of weapons research. Forge denies that this justification is convincing. However, Kemp, one of the few who have discussed the ethics of weapons research, claims that it is not only permissible to undertake weapons research for defence, and maintains that it is a *civic duty* imposed on scientists to do so (Kemp 1994). Forge and Kemp therefore adopt diametrically opposed viewpoints, and the topic can be explored by comparing and assessing their respective positions.

Kemp claims that since the state has a duty to protect its citizens, and since Hobbes many political philosophers believe that this is the *raison d’être* of the state, then those who are able to assist in this vital endeavour by designing the means for defence have a duty to do so. Note that a civic duty in this sense is not a *moral* duty, as it only applies to a part of the moral community, namely one’s fellow citizens, and hence lacks the (completely) impartial character of morality. Expressed in these terms, Forge’s view is that it is moral duty *not* to undertake weapons research. Forge does not deny that states have a right to defend themselves, or defend others from aggression, nor does he deny that weapons research can help to achieve these ends. What he does deny is that this is *all* that weapons research does, namely provide the means to resist aggression, whatever the intent behind it. Note also that Kemp’s position is compatible with weapons research being permissible while Forge’s is not. Which is correct? Kemp’s view seems to be convincing, but when weapons research is analysed in more depth, this initial reaction is seen to be wrong.

Hacking maintains that what weapons research produces in the first instance is *knowledge*, instructions for making new or improved weapons, not hardware (Hacking 1986).¹¹ Weapons research is not the same thing as weapons manufacture. A scientist who successfully discharges her civic duty does not therefore provide a *unique* array of defensive weapons for her country, but a *design*, and any institution with the requisite skills and resources could produce identical weapons if it possesses the design. For instance, the Kalashnikov, the AK-47, the most widely made weapon of all time, has been manufactured from 1947 by the Soviet Union, Russia, Poland, Hungary, East Germany, Bulgaria, Romania, China, North Korea, Yugoslavia, Iraq, Egypt and India, and variants by Finland, Israel and South Africa.¹² Mikhail

Kalashnikov invented a highly effective assault rifle, but should he be praised for discharging his civic duty or condemned for introducing into the world a terrible new means to kill? Clearly, his innovation spread beyond his own country. Successful weapons tend to become ubiquitous. Both sides in the American Civil War used ‘Minié rifles’, as did both sides in the Franco-Prussian War of 1871. By the First World War, all protagonists had machine guns, quick-firing artillery as well as reliable rifles. There was less uniformity in the armaments of the Second World War, though all sides had tanks, bombers, fighters, field artillery and so forth, and there is still less uniformity today. This trend is due to the vastly increased pace of weapons research and it is to be expected that states will strive to copy successful weapons innovations as much as they can. This may not be easy and it may not always be clear how it can be done, but the very fact that many different countries come up with weapons that are remarkable similar shows that it happens. Weapons innovation cannot be contained.

If successful weapons innovation does not lead to a unique array of weapons, perhaps this does not matter if leads to weapons that can *only* be used to defend against aggression – perhaps Kemp’s position can be resurrected along the lines that designing defensive weapons is a civic duty? The role of a defensive weapon, or weapon that is postured defensively, is to actively protect an asset, but this does not necessarily amount to defending against aggression. The work of Luttwak on the levels of strategy shows the relationship between defence and its strict ‘opposite’ offence is complex and multi-faceted.¹³ One implication of Luttwak’s analysis is that a state that would embark on a war of aggression needs a range of defensive options. It needs to be able to defend itself, its ‘homeland’, from retaliation, it needs to defend captured territory and its forces in enemy territory from counter-attack, etc. Thus, if the overall theatre or grand strategy is one of aggression, with the long-term goal of capturing territory or otherwise exerting hegemony, at ‘lower’ levels defensive tactics and even defensive operations may be necessary at certain times and places, and conversely. Fighting on the Eastern Front in the Second World War, after Hitler invaded the Soviet Union, exemplified this dynamic. At the level of grand strategy, Hitler was waging an aggressive war of conquest, while Stalin was defending. But even as early as the third month of the war, the Germans were mostly on the defensive, with their infantry holding positions while their tanks, the main instrument of the blitzkrieg operations, were being repaired and refitted.¹⁴ Thereafter, the German offensive re-commenced, stalling finally in November 1941 in front of Moscow, and the next two years the protagonists defended and attacked one another in turn until the Soviet Union began to prevail. All the weapons both sides had at their disposal were used in all forms of fighting, including tanks that were deployed in mobile defence: there were no weapons *only* used for defence. Kemp’s position therefore cannot be salvaged.

FUTURE DIRECTIONS

Forge maintains that since weapons are, by their very nature, the means to harm, then research which aims to introduce new or improved weapons into the world requires justification, and that the only possible justification is that the weapon in question will prevent more harm than would be the case were it not available.¹⁵ This kind of justification lies behind all references to defence in relation to weapons research, military spending and so forth. But this demand cannot be satisfied: weapons research produces knowledge in the form of designs which can be implemented in the future in circumstances that are unknowable – as the examples discussed above have shown - and since there is no such thing as a weapon that cannot be used to aid aggression, there can be no assurance that an episode of weapons research will produce a weapon that prevents more harm than it causes. It is precisely this assurance that is needed if

weapons research is to be justified, hence Forge argues that weapons research is always unjustified. The historical record shows that wars have become more frequent and more deadly since scientific advances began to be applied systematically to weapons research, and so the facts seem to support Forge's position. But these important issues surely need more discussion, not just from philosophers and others who reflect on weapons research, but from the scientists and other experts who actually take part. Everyone would surely hope that the very rapid pace of arms production and sales will decrease in the future and that the realisation that it is in fact wrong to design new kinds of weapons would become more apparent.

Turning finally to dual use research, how should one evaluate such an endeavour? This issue is of current concern and of ongoing discussion, but the following position will be put forward here. If research can lead to bad uses, namely to the development of new weapons, as well as good ones, how does one balance the good against the bad and what does one make of claims to the effect that if a researcher only intends the good use in a given case, then her participation is justified? The position suggested here is that the good cannot 'outweigh' the bad use and that participation is therefore not justifiable whatever the intention of the researcher. Turning to the second matter first, if an agent foresees that her actions will lead to a given outcome, then she may still be held responsible for any harmful effects of those actions even if she intended some other outcome. Not only this, but if an agent was in a position to see that her actions could have harmful effects, then she may be held responsible for these even if she did not actually foresee them – these issues are discussed in detail in Forge 2008. So when it comes to dual use, the fact that a researcher does not intend the 'bad use' which her work enables does not mean that she is not accountable for it.

When it comes to balancing good uses with bad ones, the issue involves the responsibilities of the researcher and this can be expressed as follows: How are the responsibilities of participants in dual use research to be compared to one another, in such a way that a decision can be made about whether undertaking such research is morally permissible? The position which is put forward here for future discussion is this: All researchers are obliged not to undertake work that has, or risks having, harmful consequences; this much has been agreed already. There is, however, no corresponding obligation to undertake research that has positive or beneficial outcomes. There is therefore an *asymmetry* between the responsibilities of researchers. Thus, in the context of dual use research, there is on the one hand an obligation that derives from the 'bad use, not to undertake the research, but no corresponding obligation in favour of doing the work that derives from the 'good use'. It follows that moral persons should not undertake dual use research. This position is not the last word on the question. Future work on devising a way to support the idea that there is in fact a responsibility to undertake dual use research with a view to achieving the 'good use' is one such direction.

CONCLUSION

Returning to focus again explicitly to the role of scientists, if weapons research is morally wrong and without justification, then scientists, just like everyone else, should not take part. A scientist who knowingly and deliberately takes part in an activity that is morally wrong is as blameworthy as any other participant. However, there are different ways in which scientists can be said to participate. For instance, a scientist can receive a grant from the military even though she does not intend to undertake weapons research and does not see how her work might provide a basis for weapons research. But, the military only sponsors research it thinks might be useful for its own ends, including weapons research, and this is a good reason why scientist should not accept such funding. It is also possible that a scientist knows

she is conducting research that could have military application without being sponsored by the military. This happened in the late 1930s when some members at least among the scientific community believed that current advances in nuclear physics could contribute to a weapon of mass destruction. The majority of scientists working at the time, including Hahn and Strassman, the discoverers of nuclear fission, saw themselves as ‘pure’ research scientists. Szilard sought a moratorium on publishing papers on nuclear physics in order to try to prevent such nuclear weapons research, with limited success (Rose 1998; 92). This, as well as the topic of dual use, raises complex questions about the responsibility of scientists for applications of their work that they do not intend, or even foresee, but these are questions beyond the scope of this chapter.

REFERENCES

- Acheson, D. (1969). *Present at the Creation*. Norton.
- Arrigo, J. (2000). The Ethics of Weapons Research: A Framework for Discourse between Insiders and Outsiders. *Journal of Power and Ethics*, 1.
- Bridger, S. (2015). *Scientists at War: The Ethics of Cold War Weapons Research*. Harvard University Press. doi:10.4159/9780674426085
- Forge, J. (2007). No Consolation for Kalashnikov. *Philosophy Now*, 59.
- Forge, J. (2008). *The Responsible Scientist*. Pittsburgh University Press. doi:10.2307/j.ctv12pnt8v
- Forge, J. (2010). A Note on the Definition of Dual-Use. *Science and Engineering Ethics*, 16(1), 111–118. doi:10.1007/11948-009-9159-9 PMID:19685170
- Forge, J. (2013). *Designed to Kill: The Case Against Weapons Research*. Springer. doi:10.1007/978-94-007-5736-3
- Fotion, N. (1990). *Military ethics*. Hoover Institution Press.
- Glaser, A. (2017). *Highly Enriched Uranium, Research Reactors and the Risk of Nuclear Proliferation*. American Physical Society.
- Grandy, D. (1996) Leo Szilard. University Press of America.
- Greener, W. (1910). *The Gun and Its Development*. Cassell.
- Grice, M. (2009). *On Gunnery*. Booksurge Publishing.
- Hacking, I. (1986). Weapons Research and the Form of Scientific Knowledge. *Canadian Journal Of Philosophy, Supplementary*, 1, 2. doi:10.1080/00455091.1986.10717161
- Hall, A. (2009). *Ballistics in the Seventeenth Century*. Cambridge University Press.
- Hall, B. (1997). *Weapons and Warfare in Renaissance Europe*. Johns Hopkins University Press.
- Hoddeson, L., Henriksen, P., Meade, R., & Westfall, C. (1993). *Critical Assembly*. Cambridge University Press. doi:10.1017/CBO9780511665400
- Kalashnikov, M. (2006). *The Gun that Changed the World*. Polity Press.

- Kanaher, L. (2007). *AK-47*. Wiley.
- Kemp, K. (1994). Conducting Scientific Research for the Military as a Civic Duty. In *Ethical Issues in Scientific Research*. New York: Garland.
- Liddell Hart, B. (1973). *History of the Second World War*. London: Pan.
- Luttwak, E. (1987). *Strategy: The Logic of War and Peace*. Harvard University Press.
- Manchester, W. (2003). *The Arms of Krupp*. Back Bay Books.
- Marsden, E. (1969). *Greek and Roman Artillery: Historical Development*. Oxford University Press.
- Marsden, E. (1971). *Greek and Roman Artillery: Technical Treatises*. Oxford University Press.
- Rhodes, R. (1986) *The Making of the Atomic Bomb*. Harmondsworth: Penguin.
- Rhodes, R. (1995). *Dark Sun: The Making of the Hydrogen Bomb*. Touchstone.
- Rihill, T. (2007). *The Catapult*. Pennsylvania Westholme.
- Rose, P. (1998). *Heisenberg*. California University Press.
- Stahel, D. (2009). *Operation Barbarossa and Germany's Defeat in the East*. Cambridge University Press. doi:10.1017/CBO9780511732379
- Stankiewicz, R. (2000). The Concept of Design Space. In J. Ziman (Ed.), *Technological Innovation as an Evolutionary Process*. Cambridge University Press.
- Steele, B. (1994). Muskets and Pendulums: Benjamin Robins, Leonhard Euler and the Ballistics Revolution. *Technology and Culture*, 35.
- Walter, J. (1979). *The German Rifle*. Arms and Armour Press.
- Warren, K. (1989). *Armstrongs of Elswick*. MacMillan. doi:10.1007/978-1-349-10994-4
- Weart, S., & Szilard, G. (1978). *Leo Szilard: His Version of the Facts* (Vol. 2). MIT Press. www.moralitymatters.net
- Zabicki, D. (2006). The French 75. In S. Tucker & P. Roberts (Eds.), *World War I*. ABC-CLIO.

ADDITIONAL READING

- Forge, J. (2004). The Morality of Weapons Research. *Science and Engineering Ethics*, 10. PMID:15362709
- Forge, J. (2009). Proportionality, Just War Theory and Weapons Innovation. *Science and Engineering Ethics*, 15(1), 15. doi:10.1007/11948-008-9088-z PMID:18802788
- Forge, J. (2014). The Case Against Weapons Research. *International Journal of Technoethics*, 5.
- Gert, B. (2004). *Gert Common Morality*. Oxford University Press. doi:10.1093/0195173716.001.0001
- Gert, B. (2005). *Morality: Its Nature and Justification* (Revised Edition). Oxford University Press. doi:10.1093/0195176898.001.0001

McNeill, W. (1982). *Pursuit of Power*. Chicago University Press. doi:10.7208/chicago/9780226160191.001.0001

Miller, S. (2018). *Dual Use Science and Technology and Weapons of Mass Destruction*. Springer. doi:10.1007/978-3-319-92606-3

Orend, B. (2005). *The Morality of War*. Broadview.

Walzer, M. (1977). *Just and Unjust Wars*. Allen Lane.

ENDNOTES

¹ The writer's own webpage, www.moralitymatters.net, contains some introductory material, together with more on the topic of this chapter, and also more references.

² See especially Rhodes 1986; for his discussion of Szilard, see 28-30, and for the Frisch-Peirels Memorandum see 321-325. Weart edited Szilard's papers, which were published in 1978, see Weart and Szilard 1978. Grandy has a more objective account of Szilard's contribution. Forge 2008 Chapter 8 is a short accessible overview of the Manhattan Project and the events leading up to it. For a more technical discussion, see Hodderon *et al.* 1993.

³ Rhodes' second book, Rhodes 1995, takes up where his first book ends, at the conclusion of the Second World War, and begins with a discussion of the Soviet atomic bomb project and moves on to the decision by the US to then proceed with thermonuclear weapons. Both books are important because they give background to the decisions to actually go ahead with the respective nuclear weapons research programmes, surely among the most momentous decisions of all time. For more on the decision to proceed with thermonuclear weapons research by one who was actually involved, see Acheson 1969, especially 344-349.

⁴ It was necessary to develop special steels for guns that used the new high explosives, both of which were due to advances in chemistry in the nineteenth century, see Manchester 2003. Manchester's book gives an excellent account of Krupps, the famous, or notorious, weapons manufacturer and innovator. Warren 1989 tells the story of George Armstrong, the English equivalent of Alfred Krupp. The problem of recoil, which had plagued gunners from the very beginnings of artillery, was solved at the end of the century by the French, see Zabecki 2004.

⁵ See Greener 1910: 629-632 for the two British attempts. An excellent account of Minié is in Hall 1997: 135-147. Walter 1979 is a comprehensive account of the nineteenth century German rifle innovation, especially in regard to the Mauser.

⁶ Marsden wrote two important books on this topic, which have become classics. One gives the historical development of Greek and Roman artillery, while the second reproduces four technical treatises, from the fourth century BCE to the first century CE, which give instructions for building catapults. Rihill is a more up to date, and perhaps more readable, general text on the topic, which does not always agree with Marsden.

⁷ Ammonium nitrate is strictly speaking a substance not an artefact because it does occur naturally: I am assuming here that it has been manufactured and is in that sense an artefact.

⁸ See National Academies of Sciences, Engineering and Medicine (2016) Chapter 3 for civilian research uses of HEU. As the title of this publication, *Reducing the Use of Highly Enriched Uranium in Civilian Research Reactors*, implies, there are moves to try to reduce or eliminate HEU in research reactors.

- ⁹ Sarah Bridger's recently published *Scientists at War: The Ethics of Cold War Weapons Research* has, in spite of the title, almost nothing in it about ethics. It is however an excellent account of the attitudes of many of the leading US weapons researchers between the launching of Sputnik and Star Wars, with special reference to weapons research and the Vietnam War.
- ¹⁰ Some Just War Theorists, and some others, have argued that certain classes of weapons, especially weapons of mass destruction, are immoral but have not (even) gone on to state the corollary to that proposition that the weapons research that aims to produce weapons of mass destruction is immoral. Fotion, Fotion 1990, gives one of the best discussion of the issue from this restricted viewpoint.
- ¹¹ This knowledge, in the form of designs, can be more or less well articulated, from the most basic and sketchy idea – for instance, using a chain reaction in a uranium assembly – to an engineering specification that could, in principle, be implemented entirely by machine. For more on this issue, see Stankiewicz 2000.
- ¹² For Kalashnikov's story in his own words, which must be taken with a grain of salt, see Kalashnikov 2006. For more details, and more objectivity, see Kanaher 2007.
- ¹³ Luttwak 1987 is the main source. But see also Liddell Hart 1973 for various ideas about what could count as a defensive weapon.
- ¹⁴ See Stahel 200, especially Chapter 7 and the introduction.
- ¹⁵ Forge's main work is Forge 2013, which contains many further references.

Mobilization: Decision Theory

Idongesit Williams

 <https://orcid.org/0000-0002-5398-3579>

Aalborg University, Denmark

INTRODUCTION

Every action or inaction performed by a human being is as a result of a rational process. This is a process that occurs, in some cases, in a manner of seconds. In organizations, it might take a longer time. However, whatever is the case, in the process of rationality involves the weighing of factors, making a decision and mobilizing to make that decision. Ultimately, what is being mobilized are resources. These could be tangible (including human resources) or intangible resources. However, the most common resource mobilized to perform an action is human resources. This is evident in individuals, communities and organization. However, as times change and we move into the age of automation, more machines will be mobilized than humans. Nevertheless, in this chapter, the focus will be in the mobilization of human resources in a broader sense.

The mobilization of human resources to perform defined and collective action is as old as the existence of humans. The mobilization of humans is existent in organizations, societies and families. Furthermore, mobilization is not a spontaneous act even though it might occur in seconds. Rather, there are rational factors that trigger the desire to mobilize. However, understanding what makes us humans mobilize is rarely treated as a subject in literature. There are different studies on behavioral tendencies of humans and in some of these theories, triggers to taking certain actions are identified. However, these are contextual studies and not a generalized theory that actually explains what goes into the process of organizing. As a result, managers, innovators and visionaries have to rely on their guts and inborn talent to attempt to mobilize others to perform a certain action. In most cases, because of ignorance of how to go about the process, their visions and ideas remain dormant. Hence, it is necessary to understand the factors that ought to be in place before mobilization occurs and furthermore it is necessary to understand the circumstances by which another human will submit to being mobilized without coercion. Hence, the guiding question in this chapter is when and why do we decide to mobilize? Furthermore, what are the factors that trigger the decision within us to mobilize humans to perform an action?

In responding to this question, the author will extend the work on his Ph.D. thesis by combining a set of hypothesis (models) from that thesis with social theories to theories, why we decide to mobilize. Due to the word limitation in this chapter, reference will be made to the thesis. The models from the thesis are called the Community Broadband Network Mobilization (CBNM) Models (Williams, 2016). These models, though not developed to explain mobilization, has a clue on why mobilization occurs. The theories used will be Social movement theories, Logic of collective action and the Actor network Theory. The theorization involved a deductive and an inductive process. The resulting theory was the Mobilization-decision theory. This theory provides an explanation on why humans mobilize. The theory as will be explained in the chapter is useful both to individuals, organizations, communities etc.

DOI: 10.4018/978-1-7998-3479-3.ch098

This chapter consists of 7 sections. Section 2 is a review on the theories of mobilization. Section 3 presents and explains the CBNM models. Section 4 highlights the transition from CBNM models to Mobilization decision theory. Section 5 highlights the theoretical and practical implications of the Mobilization-decision theory. Section 6 highlights the future research direction. Section 7 is the conclusion.

BACKGROUND 1 - A REVIEW THEORIES WITH CLUES TO MOBILIZATION

There are different theories that identify mobilization as a conceptual process that results in an action. Most of these social theories that explain and analyze social movements (Peterson, 1989). However, social theories that investigate the concept of mobilization are rare. As a result, questions as to why and how mobilization occurs are often left unanswered. Answers to these questions can be induced from social theories. This is because these theories explain and analyze collective behavior. These theories also provide explanations for “*why social movements emerge*” (Peterson, 1989). There are other theories one can rely on. However, in this section, the supporting theories of from which the answers will be deduced are the Actor Network Theory (ANT), and the Logic of collective behavior. The community mobilization approach was considered but later, dropped. This was because ANT and the social movement theories seemed to hold the key to explaining community mobilization. The ANT provides clues as to why and how mobilization could occur and the Logic of collective behavior because it presents a clue on people or groups allow themselves to be mobilized (Olson, 1971).

The Logic of Collective Action and Reasons for Mobilization

The logic of collective action by Olsen (Olson, 1971) is one of the theories of collective action. This theory does not explain why mobilization occurs. However, it does expose what motivates individuals to be part of a large group. The theory indicates that in the exception of the existence of an incentive or coercion, individuals in large groups serve self-interests rather than the interest of the group (ibid). In other words, if the individual is not coerced, they are part of the group because of the perceived value they will derive within the group. This value is domicile in the incentives that exist in the group.

Although Olsen argues that individualism is minimized in small groups as compared to large groups, it could still be induced that the individual agents that signed up to become a part of a small group are also driven by incentives which highlights the perceived value evident in joining the group. Hence, from this theory the role of incentives, such as perceived value is highlighted.

Social Movement Theory and Reasons for Mobilization

Social Movement theories are human-centric theories. However, although mobilization is implied in the social movement theories, reasons as to why mobilization could occur can be deduced from these theories. In order to understand the Social movement theory, one has to understand the concept of social movement. Social movement is a concept with different definitions (Mario, 1992). For example, Mario 1992 indicates that social movements can be identified as a “network of informal interactions between a plurality of individuals, groups and/or organizations, engaged in a political or cultural conflict, on the basis of a shared collective identity”. Mario’s classification, though revealing, tends to limit social movements to being political and cultural in nature. This could be that most research into social movements were on political and cultural conflicts such as the civil rights movement, the women’s right movement,

anti-colonial movement etc. (examples (Sen & Ömer, 2016) (Taylor, 1989) (Meyer & Tarrow, 1998). However, other authors also identify the economic aspects as well (See example (Sen & Ömer, 2016)).

However, social movements are triggered by social problems that resulting from dissatisfaction (see (Andrée, Clark, Levkoe, & Lowitt, 2019)). Different social theories point to different sources of dissatisfaction. These sources of dissatisfaction could hold the clue as to why mobilization takes place. A social movement theory such as the deprivation theory points to the dissatisfaction resulting from the *deprivation of societal resources* to a certain portion of the population (see example (Asingo, 2017)). Another theory, the mass-society theory point to dissatisfaction arising from the feeling of being marginalized by society (see (Berdayes, 2018) (Buechler, 2013)). Furthermore, societal changes caused by deprivation and marginalization, experienced by certain class of people in society, among other factors could result in the challenge of the social structure of the society resulting in a structural strain within that society. The structural-strain theory points to dissatisfaction arising from the discontent in the functioning and structure of an organized society (Smelser, 1963) (Langa & Benski, 2018) (Nulman & Schlembech, 2017). These dissatisfactions indicates the commonality of problems. This implies that if a group of persons identifies “common problems” that could possibly result in mobilization.

However, dissatisfaction or common problems are not the only precursors to mobilization. The social movement theories, such as the Value Added Theory, indicates that mobilization will become possible if individuals can identify the value that can be derived as being part of the group. The value oriented belief could for example be triggered by structural strain (Smelser, 1963), or perceived opportunity to advance individual interest (Weeber & Rodeheaver, 2005).

Hence perceived value is once again identified either as an incentive to mobilize or to be mobilized. Another precursor to mobilize is the potential to mobilize and utilize resources. This is explained by the Resource Mobilization Theories (McCarthy & Mayer, 1977). Another factor is on how social actors create and conceptualize the ideas either resulting in the formation or joining social movements. This is explained by the frame analysis theory (Mertala, 2017). Hence, if the person/group mobilizing and the person/group being mobilized can align in how they foresee the problems, solutions and, the perceive the organizational capacity of the mobilizer, chances of mobilization taking place is high. The alignment of the framing process between the mobilizer and the mobilized can be described as mental congruence (similarity in thought) or a convergence of rationality.

Therefore, some of the identified reasons as to why mobilization could occur include, the existence of common problems (social factor), the identification of perceived individual value (perceived value)- (value factor), mental congruence (frame related factor), and the existence and potential of the needed resources (resource related factor). However, the identified social factor is a form of frame related factor. This is because the identification of common problems is a form of diagnostic framing (Benford & Snow, 2000). Hence, the emerging factors would be the frame related factors, the value factor and the resource related factor. These factors can also be viewed as incentives, both intrinsic and extrinsic incentives, which could warrant mobilization and the willingness to be mobilized. These intrinsic factors include mental congruence, perceived value and common problems. The extrinsic factors the existence or potential of resources.

Actor Network Theory and Reasons for Mobilization

Actor Network Theory (ANT) also holds clues to the process of mobilization and being mobilized. These clues are more evident in the theorization of the theory by Michael Callon. This theorization process

by Michael Callon hold the key as to why mobilization occurs. In the ANT, mobilization occurs in two ways. The first is via the process of translation and the second is at the end of the translation process.

He theorized ANT via his observation of a group of scientist in their process of mobilizing human and non-human actors. The aim of the mobilization process was to unite the actors towards collaborating to promote a sustained production of scallops in the St. Breuic bay (Callon, 1986). The actors were the researchers, the fishermen, the scallops and scientific colleagues of the researchers that knew nothing about scallops. The researchers were the focal actors or the mobilizers. The other actors are the mobilized. In the problematization stage, the focal actors had to identify the potential resources including potential actors with which they perceive that they share common problems and perceive that, based on their plan target partners will be willing to work together with them due to mutual interest. The willingness to work together can be identified as Cooperativeness. Cooperativeness is a personality trait (see (Carriere, et al., 2019) (Gambetti & Giusberti, 2019)). The judgement with respect to cooperativeness could be either perception or evidence based. At this stage in the ANT, it is a perception. Hence “common problem” and “Perceived cooperativeness” since they indicate framing processes hence they are forms of mental congruence. However, still in the problematization stage, the focal actors (the mobilizers) had to map out a process by what each actor will do and how all the actors will work together to achieve the expected goals. The process of interposition or interressement follows this stage, as Michael Callon calls it. At this stage, the mobilizer negotiates with the identified actors towards accepting the roles assigned to them. If there is evidence of potential resources, mental congruence; individually perceived value for the mobilized; and the possibility to work together with other potential actors (cooperativeness), they will accept the invitation. Otherwise, the translation process fails. Cooperativeness at this stage is not a perception, but an evidenced based assessment. The evidence here indicates the fact that the partners are actually signing up or willing to sign up to the network. Hence, the actors are aware that the other actors are willing to work together with them. Once the actors, that are to be mobilized, accept the role assigned to them then they are said to have been enrolled. At this stage, it could be said in the literal sense that mobilization occurs at this stage. However, from the theoretical point of view, mobilization occurs at the last stage. At this stage, the mobilization of allies, is when the different actants begin to align, based on their interest and they elect one of the actants as their spokesperson in the network.

Hence, by induction, based on the four stages of translation, mobilization is possible because of the existence of potential resources, including actants, perceived value, cooperativeness and mental congruence. These factors are also intrinsic and extrinsic incentives, which are similar to those identified from the Social movement theories.

Linking the Three Theories

Based on the induction process extracted from the three theoretical approaches, a link can be identified. The link is represented in Table 1.

As seen in the table 1, what triggers motivation are incentives. These are incentives for the mobilized and the mobilizer. These incentives could be extrinsic or intrinsic in nature. The incentives identified from the ANT and the theory of social mobilization include “potential resources”, “perceived value”, “mental congruence” and “cooperativeness”. Aside, “cooperativeness”, these concepts induced from the ANT and the Social movement theories are similar.

But are these factors enough to trigger a decision to mobilize or be mobilized? In section 3 an existing model (CBNM models) that also explains why people mobilize is described. The model and the theoretical outcomes in section 2 are compared and analysed in section 4.

Table 1. The link between the three theories

Logic of collective action	Nature of the incentives	Social Movement Theories		Actor Network Theory	Main keywords
Incentive	Extrinsic incentives	Resource induced factors	Existing and potential resources	Potential resources	Potential resources
	Intrinsic incentives	Value induced factors	Perceived value	Perceived value	Perceived Value
		Frame induced factors	Mental congruence	Mental congruence	Mental congruence
				Cooperativeness	Cooperativeness

BACKGROUND 2 - ABOUT THE CBMM MODELS

The Community Based Network Mobilization (CBNM) models is a community informatics model. It emerged as one of the surprised findings in the PhD research on community mobilization. The research was aimed at finding out a suitable Public Private Interplay model for developing telecommunications networks in rural areas (see (Williams, Analysis of Public Private Interplay Frameworks in the Development of Rural Telecommunications Infrastructure: A Multiple-Case Study, 2015)). Furthermore, one of the activities in the PhD research was to identify why and how rural communities around the world develop broadband networks. Hence, the search for possibilities of implementing broadband infrastructure was the focus. A study was made of the following cases, namely the Johannesburg Wireless User Group (South Africa), Ghana Wireless network (Ghana), Dharamsala Wireless network (India), Magnolia Road Internet Coop (USA), Hallaryd Broadband Coop (Sweden), Djurslandsnet (Denmark). Data gathered from interviews, and documents from these cases were analyzed using the Grounded Theory methodology. The result of this process in that research was the emergence of the CBNM models. These models, surprisingly, highlighted the factors that led members of these communities to mobilize themselves to develop the broadband internet network.

These models had no name initially. However, as the author pondered over it for a while, it was clear that the models did provide initial insights into why community mobilization actually takes place. Hence, this model was then used to explain how people in rural areas could be mobilized to develop, 5G networks in rural areas (Williams, 2016), broadband networks (Williams, Falch, & Tadayoni, 2017), broadband networks in order to access e-government services (Williams, 2017) and later 5G Wi-Fi networks (Williams, 2018). It was in this process that the models the name of the model evolved to CBNM models.

There are four almost identical CBNM models. However, three of them are of relevance to this chapter. The three models were the developing country models, the developed country models and the universal model. The developing country models were developed from the developing country cases. These were the Indian, Ghana and South African cases. The developed country models were developed from the developed country cases. These were the Danish, Swedish and the US cases. Both models had similar constructs, but had a slight variance. Although the variance was not identified in the developing country cases, its implied implication was clear. Hence, in merging both models, the universal model was created.

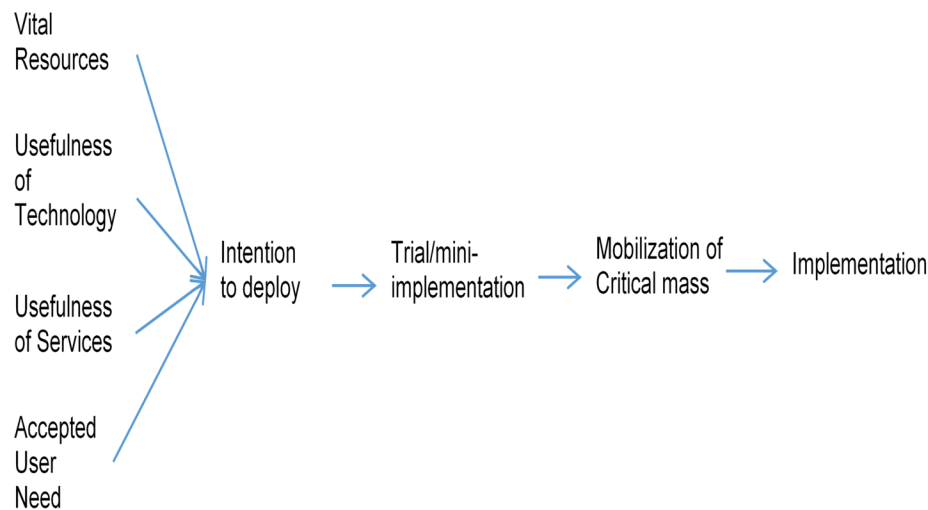
In each of the models, two classes of stakeholders can be identified. These are the visionaries (the mobilizer) and the target partners (the mobilized). The visionaries and target partners are either individuals or a group. Target stakeholders could either be innovators, early adopters, late adopters or laggards as explained in the Diffusion of Innovation Theory (Everett, 2003).

Developing Country Model

The developing country model indicates that people in rural areas will decide to mobilize in order to develop rural broadband infrastructure if they find the technology useful (usefulness of technology); find the expected service useful (Usefulness of service); identify with the need the service solves (accepted user need) and have the contextual requisite resources (vital resources) to develop the infrastructure. This is represented in the Figure 1.

Figure 1. CBNM model for Developing Countries

Source (Williams, 2015)



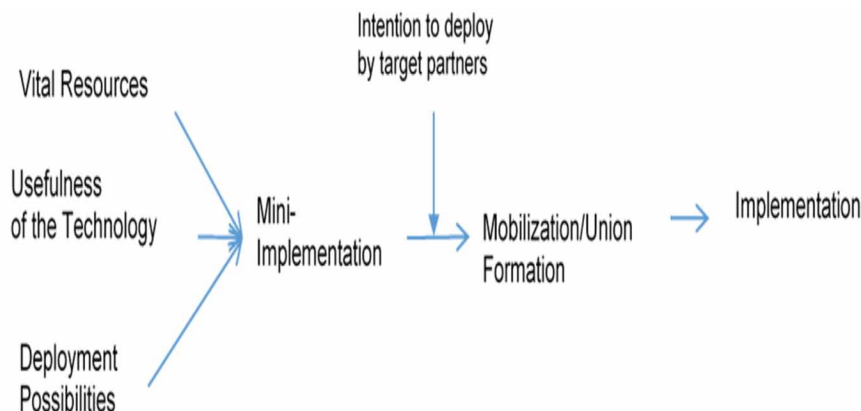
The decision to mobilize process in figure 1 involves the “intention to deploy” and the “trial/mini-implementation”. The intention to deploy is a decision first made by the visionaries. They produce a proof of concept via the trial/mini-implementation process. A successful proof-of-concept (trial/mini-implementation) reveals the four independent constructs to the target partners. Hence, a successful trial/mini-implementation reveals the feasibility of developing the initiative as a group. This triggers the “decision to deploy” in the minds of the target partners. This results in the target partners providing their consent to join or to be mobilized to facilitate the development initiative. This iteration continues until the visionaries acquire the required critical mass needed to implement the development project.

Developed Country Model

In the developed country cases, the rural residents were already aware of the usefulness of the service and were already using the service. They were not happy with the technology used in delivering the service. Therefore, the concepts “usefulness of the service” and “accepted user need” did not emerge as precursors to their decision to facilitate the development project. However since they had access to other albeit poor services, the “intention to deploy” was a mental decision made after they were sure that a particular technology lived up to its billing in the provision of expected data rates and quality of service needed. What was also important to them was the fact that there were deployment possibilities for this technology which should be within their resources.

Figure 2. CBNM Models for Developed Countries

Source (Williams, 2015) (Williams, 2016)



Therefore, the developing country model indicates that people in rural areas will mobilize themselves to develop rural broadband infrastructure if they find the technology useful (usefulness of technology); identify deployment possibilities (deployment possibilities) and have the contextual requisite resources (vital resources) to develop the infrastructure. This is represented in Figure 2.

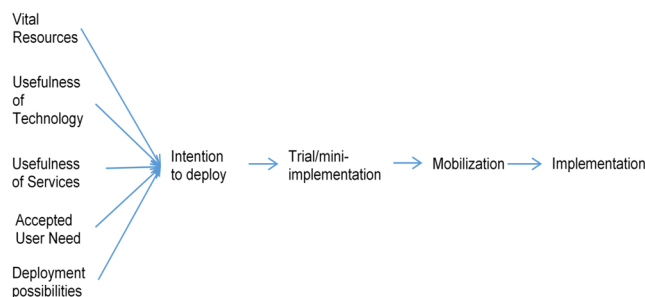
In the Danish and the US case, similar iterations and interaction between the visionaries and the target partners, as seen in the developing country model, was evident. This was not the case for the Swedish case. However, the reason and mode of iteration was different from that of the developing countries. In this case the mini-implementation was not a proof-of-concept. It was an actual infrastructure deployment by the visionaries, who initially had other plans towards the scaling of these networks. Hence, it was a self-mobilized effort. When neighbors saw the output of the initial deployments, they decided to sign up with the visionaries' in order to extend connectivity to their homes. It is in this manner that the visionaries were able to co-opt the target partners into the network.

The Universal CBNM model

The general CBNM model indicates that people in rural areas will decide to mobilize themselves to develop rural broadband infrastructure if they find the technology useful (usefulness of technology); find the service useful (Usefulness of service); identify with the need the service solves (accepted user need); have the contextual requisite resources (vital resources) and identify deployment possibilities (deployment possibilities). This model is represented in Figure 3.

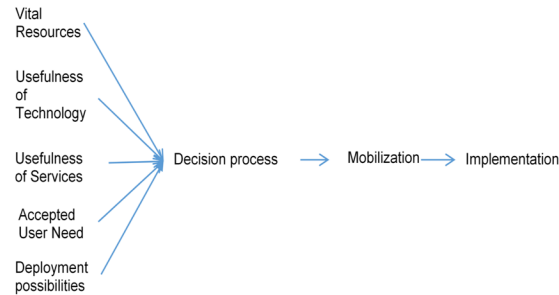
Figure 3. Universal CBNM model

Source (Williams, 2015)



The model has 4 stages, the trigger (the 5 constructs), the decision process (intention to deploy and trial/mini-implementation), mobilization and action (implementation). The general model can be re-represented as seen in Figure 4.

Figure 4. Modified CBNM model



Based on the modification of the model the decision process, even though it ultimately leads to implementation, actually moderates the likelihood for mobilization. The model though single explains two mobilization processes. The first is the self-mobilization by the visionaries be they an individual or group. The second in the factors, the visionaries present to the target partners as a selling point in order to mobilize them. The universal CBNM models also reveal the rationality pathway for the visionaries and target partners. The decision to either mobilize /be mobilized or otherwise is a product of conviction based on rationality. The rationality pathway is guided by the existence of extrinsic and intrinsic factors. The extrinsic factors is the vital resources, while the other factors are intrinsic factors. These factors are also incentives, which actually trigger the decision to mobilize and to be mobilized. However, do these universal CBNM factors align, relate or compliment the incentives identified in section 2. If there is some alignment, relationship or complementarity, then a theory that explains the decision to mobilize and be mobilized can be expressed.

FOCUS OF ARTICLE - FROM CBNM MODELS TO MOBILIZATION DECISION.

Decontextualizing the Universal CBNM Model

As the focus of this chapter is on the decision to mobilize, hence the implementation bit of the universal CBNM model becomes irrelevant. This is because mobilization could be for any reason. Hence, the universal CBNM model of Figure 5, would then be as represented in the figure 5.

The next step is to simplify the decision process to a “yes” or “no”. The reason being that the iterations in the universal CBNM model highlighted the sufficiency of the five factors towards yield to a yes answer towards mobilization in the context of study.

The next step is to find out if the concepts identified in the theory and the independent variables of the universal CBNM model match. Since the five factors of the universal CBNM models are contextualized to the cases studied, they have to be decontextualized. To do this, the identified concepts from section 2 serve as a guidepost.

Figure 5. Conversion from CBNM model to Mobilization-Decision process

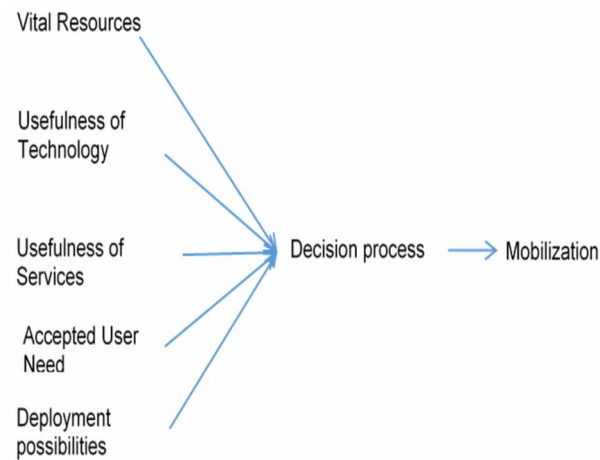


Table 2. Comparison between CBNM Independent constructs and theoretical constructs

Logic of collective action	Nature of the incentives	Social Movement theories		Actor Network Theory	Main keywords	CBNM constructs
Incentive	Extrinsic incentives	Resource induced factors	Existing and potential resources	Potential resources	Potential resources	Vital resources
	Intrinsic incentives	Value induced factors	Perceived value	Perceived value	Perceived Value	Usefulness of technology
						usefulness of service
		Frame induced factors	Mental congruence	Mental congruence	Mental congruence	Deployment possibility
						Accepted user need
				Cooperativeness	Cooperativeness	

As seen in the Table 2, there is a correlation between vital resources (CBNM) and potential resources (Keywords). However, both concepts denote the evidence or perceived evidence of the needed resources for both the visionary and the target partners. Hence, this construct can be renamed “evidence of resource”. The commonality of problems correlate with the “accepted user need”. This implies that the visionary and the target stakeholders identify the problem and that they realize and accept that they have a need. However, this factor is an aspect of mental congruence. Furthermore, there is an alignment between mental congruence and deployment possibility. This is because whatever plans are drawn as deployment possibility has to be mentally validated by the visionaries and the target partners as feasible. Hence, there has to be mental congruence before the visionary can mobilize and the target partner can agree to be mobilized. Hence, “mental congruence” is sustained as a factor. In the same vain, “usefulness of technology” and the “usefulness of service” are forms of “perceived value” first identified by the visionaries and later by the target partners. Obviously perceived value is not limited to these three constructs, perceive value could be economic, cultural, social etc. Hence, “perceived value” is chosen as a factor.

The theoretical keywords without CBNM equivalents are “common problems” and “cooperativeness”. Therefore it could be said, that the constructs in the universal CBNM models do align with some of the theoretical concepts or keywords. This would imply that the universal CBNM models on their own does explain the decision to mobilize and to be mobilized from a contextual point of view. However, it also indicates that the concepts identified in section 2 could serve as a starting point towards theorizing why the decision to mobilize occurs.

Despite this observation, how the concept “cooperativeness” does fits into the new and emerging theory. In the CBNM models, cooperativeness were not evident as a construct. However, cooperativeness was evident in the cases studied. The reason it was not captured in the data that produced the CBNM models was because the focus was not on mobilization but on implementation. However, the individuals with their selfish-ambitions would not have been mobilized if they were either mistrustful of each other or were unsure if other individuals would be cooperate. Hence, the evidence of “cooperativeness” is relevant.

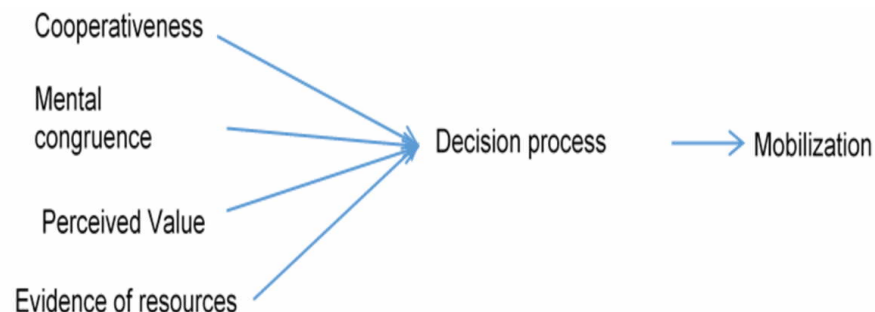
Therefore, in the generalization and comparison of the CBNM models with existing theoretical outlook, it can be deduced that the incentives for mobilization are “Evidence of resource”, “perceived value”, “mental congruence” between the mobilizer and the mobilized and “cooperativeness”. If these four conditions exist, the decision to mobilize and be mobilized will occur.

Theorizing the Decision to Mobilize

Therefore, it can be theorized that the decision process towards mobilizing a person or group of persons to perform a certain action will occur if the following conditions are satisfied. These are the evidence of needed resources, the mental congruence between the mobilizer and the mobilized; the evidence of either actual or potential cooperativeness and the perception of individual value to be derived from the action to be performed.

This theory is called the Mobilization-decision theory. The decision process is positive if the person or persons to be mobilized are actually mobilized. Hence, the perceptions of the visionary was correct. If the result is otherwise, then there was a misjudgment of the perceptions by the visionary. The diagrammatic representation of the theory is shown in the diagram below.

Figure 6. Mobilization-Decision Theory



The decision process is the factor and as mention earlier, it is either a “yes” or “no” decision. The determination of the different relevant variables, bases on social theories can be determined statistically. Furthermore, the weight determinant of the four independent variables which would indicate if the decision will be “yes” or “no” can be determined statistically.

However, if one intends to mobilize others then the four factors, supported by the explanation of each factors, in this paper can serve as a “shopping list”. This shopping list would act as a selling point to the person or group to be mobilized. If the mobilizer is satisfied with the “shopping list”, then the decision to self-mobilize is “yes”. At this point mental congruence will be an assumption. It can only be verified once the target partner grants a confirmation. In the same vain, if the mobilized is satisfied with the target partner is satisfied with the “shopping list”, and then the decision to be mobilized is “yes”. Otherwise the decision will be “No”.

Hence the mobilization-decision theory can be contextualized for the visionaries (the mobilizer) for self-mobilization and the target partner (the mobilized).

For the visionaries, it can be theorized that the decision process towards mobilizing themselves and other persons or group of persons to perform a certain action will occur if there is:

1. An evidence of needed resources;
2. Perceived mental congruence with the target partner;
3. the evidence of either actual or potential cooperativeness with and between target partners;
4. *and the perception of individual value to be derived by the visionaries and the target partners from the action to be performed.*

For the target partners, it can be theorized that the decision process towards a person or group of persons allowing himself or themselves to be mobilized to perform a certain action will occur if there is:

1. An evidence of needed resources
2. A mental congruence with the mobilizer;
3. the perception of either actual or potential cooperativeness with the mobilizer and other target partners;
4. *and the perception of individual value to be derived from the action to be performed.*

Mobilization-decision theory works for individual to group mobilization efforts and group-to-group mobilization effort.

SOLUTION AND RECOMMENDATION - THEORETICAL AND PRACTICAL IMPLICATIONS MOBILIZATION-DECISION THEORY

In principle, the mobilization-decision theory can be used to explain why people mobilize themselves and why they mobilize others. It can also be used as a framework for mobilizing people. In order to do so, theories such as Resource mobilization theory could aid in the identification and acquisition of resources. Frame analysis theories could be used as an aid to identify the approach towards identifying the worldviews required. The mental congruence could be either prognostic, diagnostic or motivational (Benford & Snow, 2000). Economic theories explaining value and value added theories of social movement can be used as an aid to identify the relevant value factors needed. Moreover, the theory of cooperation (Gardner, Griffin, & West, 2016) can be used to identify relevant factors that could serve as a frame and evidence that could serve as a basis for harnessing or identifying cooperativeness among potential target partners.

Aside its explanatory powers, the mobilization-decision theory, which borrows a great deal from the social movement theories also complements these theories. This is because the mobilization-decision theory is more explicit on the possible reasons individuals and groups decide to mobilize or be mobilized into social movements. Furthermore, the theory adds an insight in which academics could build upon or modify in order to understand why mobilization occurs and the power relations involved in the mobilization process.

Aside the theoretical implications, there are practical implications as well. The practical implication is not just in the academia but also in different types of organization. The bigger the organization, the bigger the need to create small manageable teams or task forces. In each of these teams, the visionary is often the leader of the team and this leader has to mobilize the team members to perform different tasks. Although team leaders are more likely to coerce, research indicates that coercion could result in resistance and unhealthy working environments (Truter, 2008) (Armas, De Leon, Santos, & Vitug, 2019). Hence, team leaders have to adopt innovative ways and means of getting their team members to see the need to work on a project. They have to be mobilized and be willing to be mobilized. Furthermore, in organizations, a member of a team could have an idea. At that moment, the team member is the visionary. That team member using the motivation-decision theory as a guide may successfully convince the mid-level or high-level management to embark upon the vision. Hence, the mobilization decision theory serves as a good tool.

Furthermore, still on the practical implications within the dimensions of culture, politics, education and the economy to mention a few. In each of these dimensions, there are ideas that warrant collective action. As examples, in the case of culture, the idea could be the desire to either promote or preserve cultural heritage. In the case of politics, the idea could be the case of promoting a political ideology. In the case of education, the idea could be the desire to produce more female engineers within a society. In the case of the economy, the idea could be the desire to transition from a command economic system to a market based economic system. The ideas identified in each of these examples, will only remain as ideas until a person or group of persons (visionaries) decided to act upon the ideas. These ideas require collective action. However before collective action, based on free will, can become possible then the visionary must be able to identify the perceived value of the idea, the evidence of resources for the idea, evidence of potential collectiveness among the target partners and evidence of mental congruence about the problem to be solved and the proposed solution. If the visionary is convinced by the weight of evidence emerging from these four factors, then the decision to mobilize the target partners will be made. At this point, the visionaries are mobilized to the idea. If the visionaries are able to convince the target partners of the existence of these 'evidences', then the target partners will buy into the idea, and allow themselves to be mobilized. This analogy is evident in the formation of political parties, development of government policies, think tanks, government committees in a democratic system of government, in the adoption of educational systems etc.

Apart from organizations and other practical cases mentioned, mobilization is also critical in society, not just for social movements, but also for community mobilization initiatives, such as those that produced the CBNM models. There are various areas in the world where government resources are insufficient to deliver the requisite social amenities or social programmes. If community leaders, guided by the principles of this theory could be able to facilitate cooperativeness, identify a way of getting resources, embark on a change management process to instill value and a worldview of the solution among the community members. If they succeed in this process, they could easily mobilize the community to facilitate the much-needed social amenity or social programme.

FUTURE RESEARCH DIRECTIONS

Mobilization is an ongoing activity, which will last as long as there are living organisms on earth. At every second in life, there are persons who are at the point of making a decision towards mobilization. On the other hand, there are persons who are either mobilizing, or are at a point where they are pondering on whether they should accept being mobilized or otherwise. Hence, there exist a great deal of possibilities for the application, testing, the evaluation and upgrade of the Mobilization-decision theory. The possibilities are research possibilities, which could be approached either quantitatively, qualitatively or via the mixed method approach, irrespective of the governing research paradigm and research approach. Based on this premise, future research directions can examine how the decision to either mobilize or be mobilized can be triggered or was triggered in the following context among others. These contexts could either be sociological, economic, historical, Psychologic and other aspect of the social sciences.

Another aspect for research using this theory would be to understand, if the same principles work for collective action performed by non-humans, such as machines. Obviously, the trigger here would be humans. If the theory holds, or serve as a foundation for understanding how to enable machines and other non-humans, then its applicability will be useful for engineers. Which would imply that engineers could use the theory as checklist for understanding if a machine will operate or not. This idea is far-fetched now, but it is a research opportunity.

In relation to this book, the future research direction will be to use the theory to explain the decision to mobilize in organizations. This decision could be at the lower management, mid management and upper management level of organizations. The explanatory power from the theory has the potential to expose the strength and weaknesses in the mobilization effort of either an organization or a unit. The mobilization effort could be geared towards knowledge management efforts, mobilization of capital and the mobilization of labor etc.

These are some areas for future research.

CONCLUSION

The aim of this chapter was to theories the decision to mobilize from observations made in social theories and the CBNM models. Mobilization-decision theory is a social theory; it is a mid-ranged and generalizable to a certain degree. It is a theory with a greater latitude of applicability in various aspects of life, where mobilization of people occur. It highlights the factors that will result in a decision to embark on a mobilization process. It also highlights factors that highlight conditions by which human agents will be agreeable to being mobilized for whatever agenda as free agents. Hence, it provides an initial theoretical explanation on why the decision to mobilize and be mobilized occurs. Obviously, it is foundational theory, more work is needed in the form of validation. Furthermore, it is open to modifications as evidence emerge on why people decide to mobilize.

REFERENCES

Andrée, P., Clark, J. K., Levkoe, C. Z., & Lowitt, K. (2019). *Civil Society and Social Movements in Food System Governance*. Routledge. doi:10.4324/9780429503597

Armas, K. L., De Leon, C. L., Santos, A. R., & Vitug, E. G. (2019). Workplace Bullying in Private Companies in the Philippines: Major Cause of Employee Burnout. *International Journal of Advanced Engineering Management Science*, 5(5), 327–334.

Asingo, P. O. (2017). Relative deprivation, protests and voting in Kenya. *Commonwealth and Comparative Politics*, 65–83.

Benford, R. D., & Snow, D. A. (2000). Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26(1), 611–639. doi:10.1146/annurev.soc.26.1.611

Berdayes, V. (2018). Communication Theory After the Administered Society: The “Total Market” in the Writings of the Departamento Ecuménico de Investigaciones. *International Journal of Communication*, 12, 2989–3009.

Buechler, S. M. (2013). Mass Society Theory. In D. A. Snow, D. Della- Porta, B. Klandermans, & D. McAdams (Eds.), *The Wiley- Blackwell Encyclopedia of social and political movement*. Blackwell Publishing limited. doi:10.1002/9780470674871.wbespm473

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. In L. John (Ed.), *Power, Action and Belief: A new sociology of Knowledge* (pp. 196–223). Routledge & Kegan Paul.

Carriere, C., Michel, G., Féart, C., Pellay, H., Onorato, O., Barat, P., & Thibault, H. (2019). Relationships between emotional disorders, personality dimensions, and binge eating disorder in French obese adolescents. *Archives de Pédiatrie*, 26(3), 138–144. doi:10.1016/j.arcped.2019.02.008 PMID:30898314

Everett, R. (2003). *Diffusion of Innovation* (5th ed.). Simon and Schuster.

Gambetti, E., & Giusberti, F. (2019). Personality, decision-making styles and investments. *Journal of Behavioral and Experimental Economics*, 80, 14–24. doi:10.1016/j.socec.2019.03.002

Gardner, A., Griffin, A. S., & West, S. A. (2016). Theory of Cooperation. In *ENCYCLOPEDIA OF LIFE SCIENCES* (pp. 1-8). John Wiley & Son. doi:10.1002/9780470015902.a0021910.pub2

Langa, L., & Benski, T. (2018). Global Justice Movements: Past, Present, and Future. In B. Berberoglu (Ed.), *The Palgrave Handbook of Social Movements, Revolution, and Social Transformation*. Palgrave Macmillan.

Mario, D. (1992). The concept of Social Movement. *The Sociological Review*, 40(1), 1–25. doi:10.1111/j.1467-954X.1992.tb02943.x

McCarthy, J. D., & Mayer, M. N. (1977). Resource Mobilization and Social Movements: A Partial Theory. *American Journal of Sociology*, 86(6), 1212–1241. doi:10.1086/226464

Mertala, P. (2017). Digital technologies in early Childhood education - A frame analysis of preservice teacher's perception. *Early Child Development and Care*, 189(8), 1228–1241. doi:10.1080/03004430.2017.1372756

Meyer, D. S., & Tarrow, S. (1998). *The Social Movement Society: Contentious Politics for a New Century*. Rowman & Littlefield Publishers Inc.

Nulman, E., & Schlembech, R. (2017). Advances in social movement theory since the global financial crisis. *European Journal of Social Theory*, 21(3), 376–390. doi:10.1177/1368431017714213

Olson, M. (1971). *Logic of collective action: Public goods and the theory of groups*. Harvard University Press.

Peterson, A. (1989). Review: Social Movement Theory. *Acta Sociologica*, 32(4), 419–426. doi:10.1177/000169938903200407

Sen, A., & Ömer, A. (2016). Why Social Movements Occur: Theories of Social Movement. *Bilgi Ekonomisi ve Yönetimi Dergisi*, 126 -130.

Smelser, N. (1963). *Theory of collective behaviour*. New York: Free Press of Glencoe.

Taylor, A. (1989). Social Movement Continuity: The Women's Movement in Abeyance. *American Sociological Review*, 54(5), 761–775. doi:10.2307/2117752

Truter, I. (2008). Power in the workplace ... “put your iron hand inside a velvet glove”. *SA Pharmaceutical Journal*.

Weeber, S. C., & Rodeheaver, D. G. (2005). Militias at the Millenium: A test of Smelser's Theory of collective behaviour. *he. The Sociological Quarterly*, 44(2), 181–204. doi:10.1111/j.1533-8525.2003.tb00554.x

Williams, I. (2015). *Analysis of Public Private Interplay Frameworks in the Development of Rural Telecommunications Infrastructure: A Multiple-Case Study*. Aalborg University Press.

Williams, I. (2016). Community Based Networks and 5G. *Nordic and Baltic Journal of Information and Communications Technologies*, 2016(1), 105 - 122.

Williams, I. (2017). Community Broadband network and the opportunity for e-government services. In D. B. Mehdi Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology*, Fourth Edition (pp. 3549 - 3560). Hershey, PA: IGI.

Williams, I. (2018). Community Based networks and 5G Wifi. *Ekonomiczne Problemy Usług*, 131, 321–334. doi:10.18276/epu.2018.131/2-31

Williams, I., Falch, M., & Tadayoni, R. (2017). A Public Management Framework for wireless broadband development in rural Sub-Saharan Africa. *Marketing i Zarządzanie*, 4(50), 89–116. doi:10.18276/miz.2017.50-06

KEY TERMS AND DEFINITIONS

Collective Action: A collaborative or cooperative action performed by a group of people, governed by self-interest, aimed at achieving a common vision.

Community Mobilization: The process where a group of humans participates in a collective decision-making aimed at influencing situations that affects the group.

Mobilization: The process of assembling and engagement of resources in order to perform an action.

Mobilization-Decision: This can imply either the decision to mobilize or the decision to allow oneself to become mobilized to perform an action.

Mobilization-Decision Theory: This is a theory that explains why mobilization occurs and why people decide to allow themselves to be mobilized.

Moral, Social, and Political Responsibility in the Information Age

11

Tomas Cahlik

Charles University, Prague, Czech Republic

INTRODUCTION

Information and Communication Technologies (ICTs) have penetrated during the last 20 years all human activities everywhere on the Earth. Humanity has entered into the information age, virtual reality and even virtual worlds have been created, serious discussions are led about possible impacts of Superintelligence that could be created in not so distant future.

Basic questions of moral philosophy stay as they have always been: How are we to live? What are we to be? Basic answers are, of course, that we ought to live good lives and be good persons. Humans are social beings, that is why moral, social and political philosophy overlap. Basic questions of social and political philosophy have been always concerned with informal and formal rules in social and political structures, companies and states being the most relevant for this article. Basic answer is that rules ought to be equilibrated in the sense that social structures function well and are just. Equilibrium is not petrified; it must reflect current conditions of social and political life.

What does responsibility mean? Everybody is morally responsible: she must be able to respond to questions about her life and personality; to explain what she does and why. Leaders in all social and political structures (families, schools, churches, companies, states ..) are socially and politically responsible: they are responsible both for functioning of their structures and for the latter being just.

What are just structures? This is a complicated question with many possible answers. For the purpose of this article, we stress just three important attributes of justice: people are not exploited, are not alienated and are as free as possible.

The aim of this article is:

- To specify what “living a good life” and “being a good person” could be in the information age;
- To identify some challenges and opportunities ICTs offer for good rules in social and political structures.

Having absolutely stabilized basic questions and basic answers makes the methodology of moral, social and political philosophy quite different from the methodology in sciences. In sciences, one starts with a thorough review of previous research, specifies some new and interesting research question, makes hypotheses about possible answers and bases argumentation on data. In moral, social and political philosophy, one reflects problems of the current age in mirrors that were created centuries ago and have been polished by many reflections ever since. Forms of texts are rich: dialogs, even poems, but the most used form is an essay.

DOI: 10.4018/978-1-7998-3479-3.ch099

BACKGROUND

Literature review in research articles is used for showing that the research described in the article fits into research themes that are interesting for contemporary research community. Literature review in ethical reflections is used differently, just for illustration of ideas that have been published in the area of interest and for “opening the scene”.

Looking into the Web of Science Core Collection database in May 2019 and using keywords “information technology” and (“moral responsibility” OR “social responsibility” OR “political responsibility”), in the whole timespan from 1945 to 2019 there are 108 entries obtained (from that 107 articles or conference proceedings). 20 entries were published from 1992 to 2009, 33 entries from 2010 to 2014 and 55 since 2015. This reveals not high but steady and increasing activity on the interdisciplinary border between ICTs and moral, social or political responsibility.

Looking closer into the content of those articles, following themes can be identified since the year 2010:

- Ethical questions linked with the creation and use of “big data”, including creation of agreed standards of good practice - e.g. (Rizk&Choueiri, 2006), (Light,& McGrath, 2010), (Celen, & Seferoglu, 2013), (Chen, B., & Zhang, N., 2016);
- Development of sustainable information society - e.g. (Tsai&Chen, 2013), (Busch, 2011), (Nieme la&Ikonen&Leikas&Kantola&Kulju&Tammela&Ylikauppila, 2014) in the sense of an inclusive and environmentally friendly society; application of precautionary principle in the development of ICTs (Som&Hilty& Kohler, 2009);
- Corporate social responsibility of both ICT suppliers and users – e.g. (Arevalo-Ascanio, J. G., Bayona-Trillos, R. A., & Rico-Bautista, D. W., 2015) (Tsai&Chen, 2013), (Busch, 2011), (Vaccaro&Madsen, 2009), (Chuang, S. P., & Huang, S. J., 2018), including suggestions for standards of good practice (Patrignani&Whitehouse, 2014), (Zompras, A., & Siakas, K., 2015). and how to enable consumers to push companies to behave ethically with the use of ICTs (Watts& Wyner, 2011), (Choi, S. B., Feng, Y. T., Liu, J. J., & Zhu, Q. H., 2019);
- University social responsibility (Arntzen, 2010); new teaching and learning culture based on ICTs (Stepien, 2010), (Kozhamkulova, Z. Z., Amankeldikyzy, N., & Kabaeva, D. A., 2017), (Chen, Y. F., 2016);
- Social responsibility in different ICT supported activities – e.g. bicycles’ sharing (Zhao, X. X., 2017); police service (Ramos, S., Perez-Lopez, J. A., Abreu, R., & Ieee., 2017), media (Lee, C. G., Sung, J., Kim, J. K., Jung, I. S., & Kim, K. J., 2016).

This indicates research activity that is driven by applications and can be contrasted with the research activity from the years before, that was pushed by theoretical considerations. (Lianos, 2000) e.g. starts with sociological concepts and identifies the threat that ICTs can atomize society through making development of personal trust obsolete. Lianos uses credit card as an example: one does not need to be trusted by the provider of money, the only thing that is relevant is the validity of the card. Technical norms replace moral and social norms.

Research fields may have different dynamics. It is quite usual that after the theoretical development some themes or even the whole field dissolve in applications. Nevertheless, after some time, both practitioners and applied researchers may find it useful to return to more generalizing theoretical reflection. In this article, the most general level of ethical reflection is being considered.

Moral, social and political philosophy offer a lot of valuable ideas that have been generalized from the real life problems of the whole human history. In this article, especially ideas developed by utilitarian philosophers, existential philosophers and the proponents of the “virtue ethics” can be identified. Two more areas of ethics linked with this article are “ethics of norms” and “casuistry”. Two ideologies are shortly compared in this article: Liberalism and Marxism. Readers are advised to find more detailed summaries of those three schools, two areas and two ideologies on web.

“LIVING GOOD LIFE”, “BEING A GOOD PERSON” AND “HAVING GOOD RULES” IN THE INFORMATION AGE

Issues, Controversies, Problems

Let us identify three challenges and three opportunities ICTs directly offer in the context of this article. This gives us a framework from which the discussion of economic costs and benefits stays mostly out, even if economic costs and benefits surely have moral, social and political impacts.

Specification of the **first challenge** builds upon the above described conclusion in (Lianos, 2000) that technical norms replace moral and social norms. Problem is that the network of moral and social norms has always been considered as something absolutely necessary for the identity of specific society and that the dissolution of this network means atomization and threatens the whole society.

The second challenge is linked with the current research frontier of ICTs, with virtual reality. Economic and social thinking is strongly based on utilitarianism and its basic persuasion that for having a good life pleasures must be maximized and pains minimized. ICTs allow creating a virtual reality in which the choice of pleasures is almost unlimited and in which we are able to avoid pain. The basic question here is the authenticity of such a life. It can be easy and pleasant but it loses any sense.

Above the basic utilitarian life level, another level ought to be built in which we ask if our activities are right. What is the right activity has been discussed in Western thinking for centuries since Plato and two basic outcomes are as follows:

- In a right activity, humans are used not as tools only. E.g. authors of new software ought to take into consideration not only how much they earn but if the software allows better self-realization of users, too;
- Real circumstances must be taken into consideration in following the basic target of “being better humans”. E.g. software firms must follow the profit motive as well as the moral one to stay in the business.

ICTs have not changed these basic outcomes but have surely changed the real circumstances and this opens the stage for discussion of what the right activities are in these changed circumstances. All new technologies have always done this, but what makes ICTs specific is the scale and speed of change.

One lives a good life if one realizes his or her potential in right activities. A person can develop its potential and become a virtuous person through cultivating its virtues and fighting its vices. Good people are virtuous people and there exists a positive correlation between being good people and doing right activities. So there exists a strong link between being a good person and living a good life.

The **second challenge** of ICTs lies in the possibility that people could be sucked too much into the virtual reality, enjoy pleasures there and try to avoid reality. Their lives would lose authenticity and

they would not leave the basic utilitarian life level. Impacts on both the individual and society would be analogic to impacts of drugs addiction.

There exists a positive feedback between the first and second challenge: instead of being identified with a real society existing at some geographical place, people are identified with the usually anonymous social group sharing the same virtual reality. In the process of atomization of real society, the motivation for its members to find substitutes of social ties in virtual reality increases. This closes the loop.

In this context, we could ask about the geo-political impact of having socially atomized countries in which a lot of citizens avoid reality together with socially homogenous countries where people live more in reality. Let us imagine a possibly violent conflict and ask who would probably win in such a conflict? Or let us imagine mass immigration of people from the second type of countries to a country of the first type. What would happen? It is surely not socially responsible to ignore these scenarios.

The **first opportunity** of the virtual reality ICTs have created lies in quick detection of suspect economic, political and other social activities in a real society and in facilitating the discourse about such activities. By definition, discourse is discussion in which participants are free in the sense that they are not influenced by exogenous incentives or constraints in the expression of their opinions. Virtual reality, if it is free - without censorship - and anonymous offers an optimal environment for discourses on any topic. Discourses on suspect social activities help to create socially accepted moral norms and push firms and politicians to social responsibility.

The **second opportunity** lies in the enormous educational potential of the virtual world. For Aristotle, being educated was considered to be a virtue by itself, even if not being a moral virtue. For Plato, if people just knew what were right they would behave in that way. It may be too optimistic but we know that there exists a negative correlation between educational level in a country and e.g. crime.

Instead of discussing this intuitively clear opportunity on general level let us illustrate how ICTs can be used for benchmarking and looking for best practices in ethical issues. In (Codes of Ethics Collection), there are thousands of different ethical codes. Each of them presents norms that have been accepted by some organization (firm, association, university etc.). Ethical Codes are classified according to professional categories, one of them being "Computer and Information Science". In May 2019, 116 organizations are found in this category (in September 2015, just 56), some of them with ethical codes for specific activities, some of them with different versions of the same code showing how it has been developed in time.

Any organization interested in development of its ethical norms can use this enormous source to compare the norms it already has with norms of its peers in the Codes of Ethics Collection. In an internal discourse, it can then formulate and accept the best for itself.

We can see another positive feedback, now between both opportunities: educational material easily accessible through ICTs can be used in discourses and additional knowledge created in discourses becomes easily accessible through ICTs.

The **third opportunity** lies in using ICTs for supporting the creation of virtues and the **third challenge** in suppressing the creation of vices. The lists of moral virtues and vices have been changing during the centuries since Aristotle and there is not full consent about what ought to be on the lists. The most of virtues and vices get stronger with the repetition of either good or bad activities.

Let us take the (List of virtues) and (List of vices) from the webpage <https://www.virtuescience.com/>, choose some of them and analyze how they can be strengthened by ICTs. (We could surely take any other list, but the advantage of these two lists is that the list of virtues has much more items than the list of vices. This supports the optimistic attitude towards human nature.)

“Compassion” and “Charity”: ICTs allow watching repeatedly and with high frequency poor people around the World what develops compassion among the most of watchers; psychologically especially among young people who are in the same time the most knowledgeable users of ICTs. This positive correlation can make this effect quite strong. ICTs offer a very easy way (e.g. donation through SMS) to do charitable activities what makes compassionated people to do such activities repeatedly. This strengthens the virtue charity.

“Curiosity”, “Commitment” and “Creativity”: Anderson in (Anderson, 2010) describes a nice example how web video powers global innovation. A slum boy watches video with street dance. Watching wakes up his curiosity and he makes commitment to master street dance himself. He surely has to be highly creative in transferring the knowledge from video into reality. There are millions of different activities and millions of web users, many of them behaving similarly as the slum boy, so ICTs have a very strong effect on these three virtues.

“Chastity” as a virtue or “Lack of Chastity” as a vice: In this context, pornographic content of web sites ought to be discussed. Watching pornography repeatedly creates habit - sometimes addiction – and weakens chastity. This discussion is not relevant for all World societies, e.g. in Japan people do not consider chastity as having sexual context. Watching pornography is clearly linked with the second challenge of ICTs discussed above; real sexual activity can shift to virtual reality what could have demographic impacts.

“Sloth”: Let us understand laziness as aversion to be active in the real world. Being active in that part of virtual reality that overlaps with reality perhaps weakens sloth; ICTs can increase the efficiency of our real activity what increases our motivation for doing it. Problem is that ICTs have created an enormous potential for procrastination – for doing less urgent but pleasurable tasks in preference to more urgent but less pleasurable ones. The seductive power can be strong and yielding to temptation is surely positively correlated with sloth.

“Enthusiasm” and “Cooperation”: For economists, it is really difficult to explain the very existence of the open source software or Wikipedia because of non-existence of material incentives. Possible explanation is following: Knowledgeable users of ICTs are usually quite enthusiastic about the possibilities ICTs offer. They form a global community and identify strongly with this community. This has created the situation in which they like to cooperate not for material benefit but for achieving a shared goal. This works if individual costs of cooperation are negligible, what they with the use of ICTs can really be. Social benefits are enormous.

“Cooperation” is linked with the problem of coordination. There has always existed a trade-off between order in social structures and personal liberty. Social structures have their functions and coordinate individuals according to their functions. They put constraints on the behavior of individuals; the latter pay a “liberty tax” and get material, social and political security in exchange.

One of basic structures are companies. Since the 19th century, there exists a controversy between Marxists and liberals about justice of capitalist companies. Marxists stress the exploitation of workers by capitalists and the alienation of labor; workers are just tools for the creation of profit, wages are costs as any other costs and workers are cut-off from the results of their work. Marxist solution is socialist revolution, dictatorship of proletariat, nationalization of private property and bureaucratic coordination. Liberals stress the loss of liberty linked with Marxist solution and worse functioning of socialist economies compared with capitalist economies.

ICTs have potential to make this conflict less severe. The existence of platforms supports self-employment, where problems of exploitation and alienation do not exist. Airbnb has created millions of self-employed hoteliers, Uber millions of self-employed taxi drivers. Networks of self-employed

professionals linked with cell phones can deliver complex reconstruction services in housing in such a quality as traditionally organized construction firms. And there are many other examples.

All this works because ICTs have brought a new dimension into the creation of trust, the basic condition for successful cooperation. Before: trust was based either on personal links in small local communities or on the authority of state and big companies. Today: giving “likes” or other form of evaluation to services offered on web has this trust-creating effect.

It is understandable that traditional companies do not like this; it is a new type of competition for them. They use very often their lobbying power to try to forbid these platform based services. Responsible political reaction is surely not to forbid these services, but to find ways how to harmonize the new opportunities with the traditional thinking.

Very interesting question is linked with crypto-currencies. Their existence undermines the monopoly of the state to issue fiat money. Acceptance of fiat money is based on trust enforced by the state. Blockchain as the basic technology crypto-currencies are based on has a trust-creating potential that can replace the trust guaranteed by the state. Again, responsible political reaction is to find ways how to harmonize traditional forms of currency with crypto-currencies. It can be difficult for politicians to be responsible in this area because it undermines one of the economic reasons for the existence of the state.

Solutions and Recommendations

The positive impact of above discussed opportunities can neutralize or outweigh the negative impact of above discussed challenges - threats. We cannot say for sure what happens because we know that all systems with positive feedbacks can considerably and discontinuously change very quickly their behavior and that it is impossible to predict when the change comes and what the behavior of the system would be after the change. Similar situation has occurred with the spread of each new technology and some recommendations can be done based on historical experiences.

It would not have any sense to try to stop the development of ICTs, it is simply impossible. What is possible is to try to increase the costs of ICTs activities that are negative from the ethical point of view and to increase benefits of activities that are positive from the ethical point of view. If it ought to be done, how to do it and who ought to do it opens a lot of questions. Let us take as an example the possibility of censorship.

Censorship can decrease negative impact through forbidding access to pages that could create addiction. Problem is that it constrains freedom in the virtual reality that is the basic condition for using up the opportunities of ICTs. In each society, some equilibrium level of censorship has been created that reflects different opinions on this issue. Both opinions and the equilibrium shift in time, we can imagine that in Western societies more www pages than child pornography could be blocked and in China some of the currently blocked www pages could be unblocked.

Previous example shows that discourse of ethical issues linked with ICTs can never stop and that all members of society have responsibility to take part in it.

FUTURE RESEARCH DIRECTIONS

Following research questions would deserve more attention:

- What is the impact of ICTs on social identity? Could possible changes in social identity have a geo-political impact?
- What other virtues and vices are influenced by ICTs and how?
- Is it possible to increase the costs of ICTs activities that are negative from the ethical point of view, ought it to be done and by whom?
- Is it possible to increase the benefits of ICTs activities that are positive from the ethical point of view, ought it to be done and by whom?

CONCLUSION

ICTs offer a lot of possibilities and present some challenges – threats – to everybody for living a good life in the ethical sense and being a good person. To leaders in different social structures – companies and states being the most relevant for this article, ICTs present possibilities for the development of more efficient and more just rules.

It is impossible to predict if the positive impacts outweigh the negative ones. Anyway, it is impossible to stop the development of ICTs. The question is if the society ought to try to increase the costs of ICTs activities that are negative from the ethical point of view and to increase benefits of activities that are positive from the ethical point of view, who ought to do it and how. All members of society, but especially its leaders have responsibility to participate in discourse about this question.

REFERENCES

- Anderson, Ch. (2010). *How web video powers global innovation*. Retrieved May 14, 2019 from https://www.ted.com/talks/chris_anderson_how_web_video_powers_global_innovation
- Arevalo-Ascanio, J. G., Bayona-Trillos, R. A., & Rico-Bautista, D. W. (2015). Corporate social responsibility and innovation: A look from the information technology and communication in organizations. *Clio America*, 9(18), 180–189.
- Arntzen, A. A. B. (2010). Knowledge, Technology and University Social Responsibility (USR): A Conceptual Framework. *Proceedings of the 11th European Conference on Knowledge Management*, 1-2, 74-81.
- Botsman, R. (2016). *We have stopped trusting institutions and started trusting persons*. Retrieved May 14, 2019 from https://www.ted.com/talks/rachel_botsman_we_ve_stopped_trusting_institutions_and_started_trusting_strangers
- Busch, T. (2011). Capabilities in, capabilities out: Overcoming digital divides by promoting corporate citizenship and fair ICT. *Ethics and Information Technology*, 13(4), 339–353. doi:10.1007/10676-010-9261-3
- Celen, F. K., & Seferoglu, S. S. (2013). Investigation of Elementary School Students' Opinions Related to Unethical Behavior in the use of Information and Communication Technologies. *2nd World Conference on Educational Technology Research*, 83, 417-421.

Chen, B., & Zhang, N. (2016). Evaluating the Corporate Social Responsibility and Organizational Performance based on Large-Scale Data Access and Management. *International Journal of Grid and Distributed Computing*, 9(1), 155–165. doi:10.14257/ijgdc.2016.9.1.16

Chen, Y. F. (2016). University Teachers' Information Literacy and Its Training under the Background of Education Information. In L. Chang, C. Guiran, & L. Zhen (Eds.), *Proceedings of the International Conference on Electronics, Mechanics, Culture and Medicine* (Vol. 45, pp. 193-198). 10.2991/emcm-15.2016.35

Choi, S. B., Feng, Y. T., Liu, J. J., & Zhu, Q. H. (2019). Motivating corporate social responsibility practices under customer pressure among small- and medium-sized suppliers in China: The role of dynamic capabilities. *Corporate Social Responsibility and Environmental Management*, 26(1), 213–226. doi:10.1002/csr.1673

Chuang, S. P., & Huang, S. J. (2018). The Effect of Environmental Corporate Social Responsibility on Environmental Performance and Business Competitiveness: The Mediation of Green Information Technology Capital. *Journal of Business Ethics*, 150(4), 991–1009. doi:10.1007/10551-016-3167-x

Codes of Ethics Collection. (n.d.). Retrieved May 14, 2019 from <http://ethics.iit.edu/ecodes/introduction>

Kozhamkulova, Z. Z., Amankeldikyzy, N., & Kabaeva, D. A. (2017). Information technology used in the preparation of future teachers and their development. *News of the National Academy of Sciences of the Republic of Kazakhstan-Series Physico-Mathematical*, 2(312), 110–115.

Lee, C. G., Sung, J., Kim, J. K., Jung, I. S., & Kim, K. J. (2016). Corporate social responsibility of the media: Instrument development and validation. *Information Development*, 32(3), 554–565. doi:10.1177/0266666914559856

Lianos, M. (2000). Dangerization and the end of deviance - The institutional environment. *British Journal of Criminology*, 40(2), 261–278. doi:10.1093/bjc/40.2.261

Light, B., & McGrath, K. (2010). Ethics and social networking sites: A disclosive analysis of Facebook. *Information Technology & People*, 23(4), 290–311. doi:10.1108/09593841011087770

List of Virtues and List of Vices. (n.d.). Retrieved May 14, 2019 from <https://www.virtuescience.com/vicelist.html>

Niemela, M., Ikonen, V., Leikas, J., Kantola, K., Kulju, M., Tammela, A., & Ylikauppila, M. (2014). Human-Driven Design: A Human-Driven Approach to the Design of Technology. *Ict and Society*, 431, 78–91. doi:10.1007/978-3-662-44208-1_8

Patrignani, N., & Whitehouse, D. (2014). Slow Tech: The Bridge between Computer Ethics and Business Ethics. In K. Kimppa, D. Whitehouse, T. Kuusela, & J. Phahlamohlaka (Eds.), *Ict and Society* (Vol. 431, pp. 92–106). doi:10.1007/978-3-662-44208-1_9

Ramos, S., Perez-Lopez, J. A., Abreu, R., & IEEE. (2017). Information and Communication Systems and Technologies in the Police Force. In *2017 12th Iberian Conference on Information Systems and Technologies*. IEEE.

Rizk, N. J., Choueiri, E. M., & INSTICC. (2006). The concept of ethics in electronic qualitative research. *ICEIS 2006: Proceedings of the Eighth International Conference on Enterprise Information Systems: Information Systems analysis And Specification*, 126-134.

Som, C., Hilty, L. M., & Kohler, A. R. (2009). The Precautionary Principle as a Framework for a Sustainable Information Society. *Journal of Business Ethics*, 85(S3), 493–505. doi:10.1007/10551-009-0214-x

Stepien, T. (2010). Transcultural and transdisciplinary competences. technological transformations of society and education. *3rd International Conference of Education, Research and Innovation (Iceri2010)*, 4815-4824.

Tsai, H. Y., & Chen, W. C. (2013). Exploring Intelligent Information Communication Services: A Case of Corporate Social Responsibility. In X. H. Liu, K. F. Zhang, & M. Z. Li (Eds.), *Manufacturing Process and Equipment, Pts 1-4* (Vol. 694-697, pp. 3636–3641). doi:10.4028/www.scientific.net/AMR.694-697.3636

Vaccaro, A., & Madsen, P. (2009). Corporate dynamic transparency: The new ICT-driven ethics? *Ethics and Information Technology*, 11(2), 113–122. doi:10.1007/10676-009-9190-1

Warburg, B. (2016). *How the blockchain will radically change the economy*. Retrieved May 14, 2019 from https://www.ted.com/talks/bettina_warburg_how_the_blockchain_will_radically_transform_the_economy

Watts, S., & Wyner, G. (2011). Designing and theorizing the adoption of mobile technology-mediated ethical consumption tools. *Information Technology & People*, 24(3), 257–280. doi:10.1108/09593841111158374

Zhao, X. X. (2017). Research on current situation and problems of shared bicycle from the perspective of shared economy. In Z. Yang (Ed.), *Proceedings of 2017 China Marketing International Conference: Marketing Strategy in the Sharing Economy: Localization and Globalization* (pp. 1204-1209). Academic Press.

Zompras, A., & Siakas, K. (2015). An Investigation of ISO 26000 and Social Responsibility Practices Applied in IT Companies. *International Journal of Human Capital and Information Technology Professionals*, 6(2), 33–48. doi:10.4018/IJHCITP.2015040103

ADDITIONAL READING

Chalmers, A. F. (1999). *What is this Thing Called Science?* Univ. of Queensland Press.

KEY TERMS AND DEFINITIONS

Blockchain: In informatics, a special type of a distributed decentralized database with growing number of records that are linked with the use of cryptography. It is resistant to any modifications.

Casuistry: Branch of ethics that uses case studies to make clear moral principles and norms in typical situations. All ethical codes specify some norms of conduct. These norms are usually explained with the use of different case studies.

Crypto-Currencies: Digital assets that are based on blockchain (or other advanced information technology) and can fulfill similar functions as money.

Ethics: Practical philosophy, scientific analysis of moral contents.

Ethics of Norms: Branch of ethics that analyses the structure of norms and what they are based on.

Existential Philosophy: Branch of philosophy that starts with individual existence and the problem of its being in the World. Important predecessor was Søren Kierkegaard (19th century), among main representatives we can find Martin Heidegger and Jean-Paul Sartre (both 20th century). They looked for authentic existence and have found different solutions depending especially on the religiosity of the specific philosopher.

Liberalism: It is (not only) an ideology aiming at personal liberty, rule of law and consent of the governed.

Marxism: It is (not only) an ideology aimed at equality and building of a classless society.

Norms: Norms are the shared and sanctioned rules. Sanctions can be both formal (e.g. law) and informal (e.g. pressure of social group).

Utilitarianism: Branch of ethics that bases moral reasoning on consequences of actions. Actions are good if they maximize pleasure and minimize pain. Among grounders are Jeremy Bentham and John Stuart Mill (18th-19th century). It has penetrated the mainstream economics. It can be shown that total utility of consumers can be increased by redistribution.

Virtual Reality: Created and accessed by ICTs. It usually overlaps with “reality,” but the overlapping is changing and fuzzy. Extremes of Virtual Reality are Virtual Worlds, in which “avatars” of humans can live parallel lives.

Virtue Ethics: Branch of ethics that considers humans as being bearers of given or developed virtues. Humans have to develop virtues and fight with vices for becoming good. Basic concepts go back as far as to Plato and Aristotle (5th, 4th century BCE).

The Voice of Social Media, 1997–2018

11

Bruce Lawrence Cook

Cook Communication, USA

INTRODUCTION

Today's social media constitutes communication portals where anyone can sign up and freely "post" messages to self-selected groups of individuals. It includes three key elements: (1) a public or semipublic profile within a bounded system, (2) a list of other users with whom they share a connection, and (3) the ability to view and traverse their list of connections and those made by other users in the system¹. (Che & Ip, 2017)

Social media can be seen as ground-level communication as opposed to the alternative - mass communication. Within some ill-defined limits, individuals can post anything they want on Social Media, while communication on mass media is heavily edited by gatekeepers – those who create, develop and edit public communication. As early as 1975 Janowitz (1975) presented the gatekeeper role in journalism as a method for reporting information with scientific method for objective and valid results. In the same article, however, he compared the journalist's gatekeeper role with that of being an advocate. Surprisingly, this contrast was to reappear decades years later in communication via social media.

While most individuals haven't sought to be journalists, freedom of expression naturally accompanies the desire for individuals to publish their own ideas. Thus, when social media arrived, many individuals joined the social media movement.

Social Media as Mass Communication by Individuals

The concept of "mass communication" has traditionally represented a one-way flow from source to recipient. Early writings in communication theory also featured feedback and context, but mass communication remained a broadcast to the masses (Schramm, 1954). Before examining its properties, one might consider the state of communication late in the 20th century.

In the last decade of the century, Berger (1995, p. 7-9) distinguished between individualism and communication to large population groups. At that time, individuals could hope to communicate with persons outside their acquaintance, but it was difficult to "reach the masses". At mid-decade the new Internet had developed from small networks to the extent that it was seen as a mass medium of communication (Morris & Ogan, 1996). However, even with the Internet, the individual's ability to access the web to reach the public was limited by cost and technology.

Throughout these formative years, the main concept in mass communication was to broadcast news and entertainment into homes from a central place. News editors served as gatekeepers, automatically limiting news content. Their news stories were received in homes and, as long as the number of networks was limited, viewers, readers and listeners could discuss the shows with friends and family the next day.

When the Internet arrived, unpublished writers hoped to gain access there. In reality, it was difficult and expensive to start a website that would ever attract a large public audience like that of radio and television

DOI: 10.4018/978-1-7998-3479-3.ch100

broadcasting. And, quite naturally, those websites with many visitors simply basked in their monopoly status and continued limiting access to the Internet in the same way radio and television always had.

In retrospect, the main question must have been— who was permitted to originate news content? The average person could never hope to be “on the show”. In fact, it was clear that only the media managers, producers, and directors could decide. If they said your news wasn’t important, it wasn’t important. While many people had a desire to be “on the air” to share their ideas, this was impossible for almost everyone. Access for the average person was denied.

Background

Individuals were able to publish their own content starting in the 1980’s with the public Bulletin Board Services (BBS). The BBS was sort of a hobby medium on the early Internet service. (Kayingo & Hass, 2017) Here an individual could locate a “bulletin board” covering a specific area of interest and communicate with other individuals who had made the same choice.

The first Bulletin Board System (BBS) was created by two hobbyists in Chicago who wanted a way to inform friends about meetings, post announcements, and share interesting information. (Langmia, Tyree, O’Brien, & Sturgis, 2013).

As such, today’s social media model did not emerge until 1997 with the advent of the first social site sixdegrees.com. (Boyd & Ellison, 2007). Some web users started first “weblogs” in the same year (Oxford English Dictionary, 2015) and the word “social media” first appeared (Bercovici, 2010). Another service, ChinaRen Alumni was created in 1999, as China’s first network community that matched the social network model.

How did writers gain access? In 1999 a new trend for access emerged with a small website called Author-me.com. Its purpose, according to the Internet Archive, was as follows:

We nurture new writers who pursue a dream of publishing their work. We offer a chance to protect each manuscript, web-publish it, and receive private comments on it via your e-mail address. Here’s a bit of encouragement during the tough process of refining your writing skills and submitting manuscripts for publication.

As social platforms became established and were adopted by many users, services such as LinkedIn (started in May 2003) began to transition from a ‘web of things’ to a ‘web of people’.

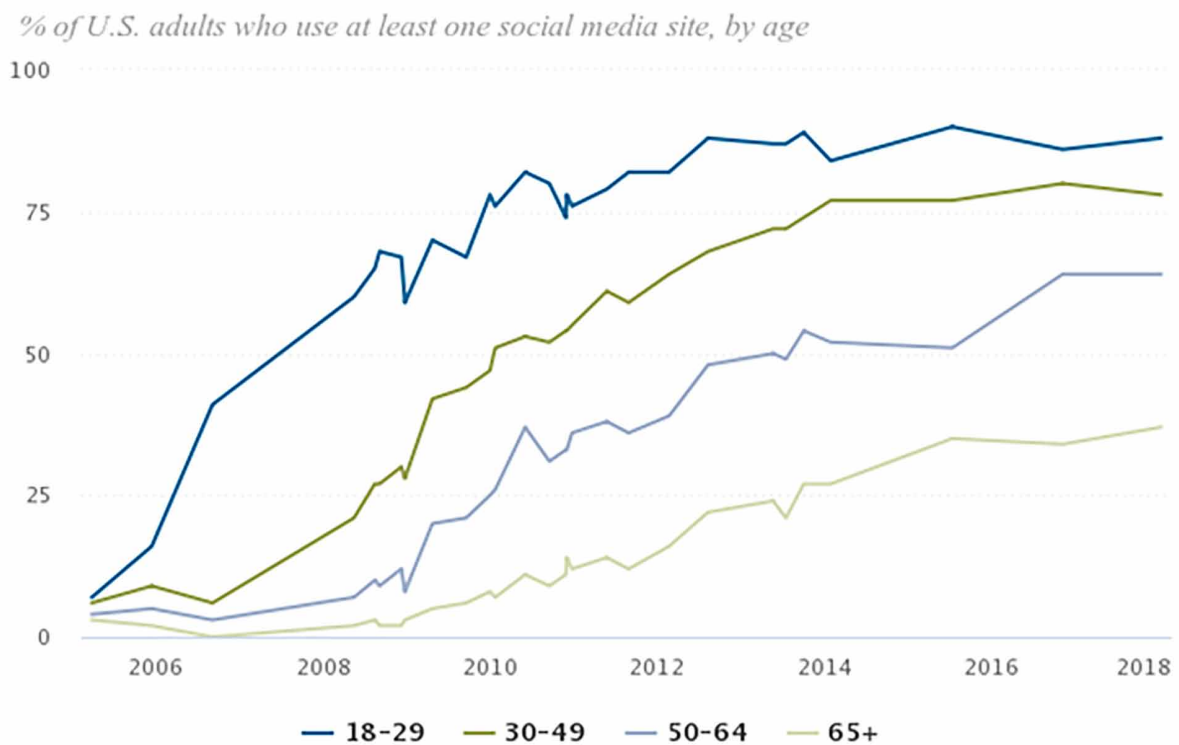
It was not until the launch of MySpace (2003), Facebook (2004) and Twitter (2006) that social networks gained widespread popularity. (Schoop, 2017). Soon these same websites converted to a new concept. Free Internet participation was available to anyone, and the social media bubble burst. This radical development began with very little fanfare.

Facebook was established by Mark Zuckerberg at Harvard University. At first, the website simply offered commercial database services on about.com. At about the same time, alternative social media websites MySpace (founded by Chris DeWolfe and Tom Anderson) gave access, along with Twitter. By 2006 Jack Dorsey, Noah Glass, Evan Williams and Biz Stone were discussing the idea of using text messaging to share “statuses”, etc., which form the basis of today’s postings. (Furlong, 2013).

As social media developed, for commercial reasons, it became important to “track” the number and location of viewers of specific pages. (Xiang & Fesenmaier, 2016). Further, users could publish “profiles” showing their background and interests. (Eastin, 2010).

The Pew Research Center first started tracking social media platform adoption in 2012 (Remland, 2016). By 2018 they could summarize: “When Pew Research Center began tracking social media adoption in 2005, just 5% of American adults used at least one of these platforms. By 2011 that share had risen to half of all Americans, and today 69% of the public uses some type of social media.” Further analysis revealed that: “As more Americans have adopted social media, the social media user base has also grown more representative of the broader population. Young adults were among the earliest social media adopters and continue to use these sites at high levels, but usage by older adults has increased in recent years.” (Pew, 2018)

Figure 1.



Source: Surveys conducted 2005–2018.
PEW RESEARCH CENTER

Multi-level Communication

Unlike journalism, the social media phenomenon permits a vast number of inputs to communication. Any user can post a profile or “status,” and any interested viewer can comment back in public or private messages. Many users can employ instantaneous reporting of a new thought, issue, or an incident, and in many cases an eyewitness account as well.

Due to the personal origin of these messages, and because any individual can post messages, social media sought a high level of credibility. The message reflects ideas of individuals, and the communication is less vulnerable to the advocacy foreseen by Janowitz in 1975 - political and economic pressures

which drive news on the broadcast media. Thus, social media can supplement main media and reflect a broader scope.

Social media also has an important role in free speech. In cases where government censorship blocks legitimate news, as has been the case in Iran, social media can provide an eyewitness account (Rahimi, 2011). This enables individuals to undermine censorship by repressive governments and helps to inhibit false propaganda. Since the reports are immediate, there's much less chance of bias in reporting. In cases when such reporting is extreme, it's possible for other witnesses to contest and correct the original posting. This is the self-correcting feature of social media.

Erik Qualman's book *Socialnomics* (2013) makes this comment:

The good news is that in this new world order, once you have determined your initial messaging strategy, you have the ability to tweak it for relevancy based on feedback....

Extreme viewpoints can be damaging to productive dialogue, but on social media it's typical for extremist views to be followed by contrary comments, leading to more useful conclusions.

The most popular postings are called "memes," and they are said to "go viral" because thousands or millions of communicators choose to "share" them with their friends and contacts.

Ideally, the development of social media led to an expectation that communication would be personal, often a conversation among acquaintances or like-minded individuals. Surely this would characterize early postings on Facebook and other social media. For these reasons, social media offers untapped potential to build harmony into international communication.

Social Media as an Instrument for World Harmony

This new medium of social media may well represent the most effective venue for building peace in today's troubled world. Occasional news stories cover evil actions in today's world, but social media can quickly assemble members into groups that mobilize and help to shift world opinion from fear to action. And, for perhaps the first time in world history, communication – not weapons – can turn the tide toward peace and harmony.

A recent "snapshot" of scholarly literature on social media and peacemaking (Cook, 2018) found considerable hope in several findings:

- 1) Social Media Supports Clustering of Citizen Groups
- 2) Social Media Influences International Relations and Activism
- 3) Social Media Communicates in Crises

In his book *Tweets and the Streets*, Paolo Gerbaudo (2012) examined the relationship between the rise of social media and the emergence of new forms of protest from the Arab Spring to the 'indignados' protests in Spain and the Occupy movement, thus illustrating the power of this medium.

From these findings it is clear that social media can quickly (albeit temporarily) mobilize support for peacebuilding efforts. While it's important for peacemakers to seek negotiated solutions, develop research reports and theories, and create artistic and educational expressions, it's important to give new attention to populist efforts to solve long-standing political rivalries and entrenched political interests.

Social Media as Activist Communicator

Social media has taken on an advocacy role which is beginning to rival that of the communication networks. While Mark Zuckerberg was a founder and developer of social media, later he had to apologize because his company shared user data with Cambridge Analytica, a company which is reputed to have used the data to affect the outcome of the US 2016 Presidential Election (Persily, 2017).

Even more surprising in the deliberate use of the social media network Twitter by Donald Trump, the President of the United States, to use daily “tweets” to communicate his personal view of political events to millions of followers (Lester, 2018), thus bypassing the gatekeeper and advocacy roles of journalists who do not share his personal views. Kellner (2018) characterizes this phenomenon as “Donald Trump and the War on the Media.” Further, similar activities by world leaders are occurring with other world leaders: Narendra Modi, Nigel Farage, and Geert Wilders (Thawani, van der Vlugt, Out, & Chandra, 2018).

Such activities have not escaped the attention of broadcast media and political parties, and social media is no longer simply seen as an alternative to dictatorial governments. However, as in experience with repressive governments, serious attention is being given to censoring content in social media. This is quite a break from the freedom of communication which first characterized social media platforms. As of 2018, Pew Research found that 72% of a sample of US adults thought social media platforms intentionally censor political viewpoints (Moon, 2018). While these controls are socially responsible, they may also constitute censorship or advocacy.

A 2018 editorial in the New York Times on suppression of “hate speech” had this comment: “If we allowed these voices to be silenced on grounds that they promote hate, we’d find ourselves scrambling to defend the radical poets, musicians, filmmakers and other artists who have pushed the boundaries of expression into what could arguably amount to hate speech, but who have done so from the vanguard of social and political protest.” (Nielsen, 2018).

Regardless, Facebook deleted posts by individuals and groups which seek to publish highly questionable, inaccurate, or inflammatory content. (Ng, 2018), (Keller, 2018). While this may seem a temporary reaction to criticism, social responsibility in content is becoming a necessity (see Weaver, 2018).

Further Research Directions

The Internet and social media radically broadened the scope of mass communication research beyond radio, television, and newspapers. While it has been common to lament errors and bias in news coverage, the introduction of social media transformed the picture by offering free speech in public to everyone. While this would seem to satisfy the critics, recent problems with social media have made it the new monster in the room. Now social media platforms stand accused of political bias (Moon, 2018) and journalism advocacy takes on a new focus. Given the relatively recent emergence of social media and abundant data on traffic and content, researchers might well address the contrast between gatekeeper and advocacy roles in mass communication.

CONCLUSION

Seen in the context of mass media and journalism, social media has become another bump in the road of unexpected competitors for traditional news of the later 20th century. While limitations of technology kept a focus on radio, television, and newspapers in the early years, the advent of cable TV, the Internet,

and social media has complicated the picture for both journalists and the audience. The shifting political and entertainment landscape presents a unique challenge for all participants, and one can only hope that social responsibility will prevail.

REFERENCES

- Bercovici, J. (2010). *Who coined “social media”? Web pioneers compete for credit*. Forbes. Disponible.
- Berger, A. A. (1995). *Essentials of mass communication theory*. Sage (Atlanta, Ga.).
- Boyd, D. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 11.
- Che, X., & Ip, B. (2017). *Social Networks in China*. Chandos Publishing.
- Cook, B. L. (2018). Social Media and Peacemaking: A Snapshot of Literature. In B. L. Cook (Ed.), *Handbook of Research on Examining Global Peacemaking in the Digital Age*. IGI Global. doi:10.4018/978-1-5225-3032-9.ch009
- Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). Demographics of key social networking platforms. Pew Research Center.
- Eastin, M. S. (Ed.). (2010). *Handbook of Research on Digital Media and Advertising: User Generated Content Consumption: User Generated Content Consumption*. IGI Global.
- Furlong, M. (2013). Putting a Price on Friendship: Examining the Ownership Battle between a Business' Social Media Networks, and the Humans That Operate Them. *The John Marshall Law Review*, 47(2), 10.
- Gerbaudo, P. (2018). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.
- Janowitz, M. (1975). Professional models in journalism: The gatekeeper and the advocate. *The Journalism Quarterly*, 52(4), 618–626. doi:10.1177/107769907505200402
- Kayingo, G., & Hass, V. M. (Eds.). (2017). *The Health Professions Educator: A Practical Guide for New and Established Faculty*. Springer Publishing Company.
- Kellner, D. (2018). Donald Trump and the War on the Media: From Election '16 into the Trump Presidency. In *The Trump Presidency, Journalism, and Democracy* (pp. 19-38). Routledge.
- Ketler, A. (2018). Facebook, Google and YouTube Censorship at an all-time high—what ever happened to Free Speech. *Collective Evolution*, 1.
- Langmia, K., Tyree, T. C., O'Brien, P., & Sturgis, I. (Eds.). (2013). *Social media: pedagogy and practice*. University Press of America.
- Lester, P. M. (2018). Social Media. In *Visual Ethics* (pp. 117–128). Focal Press.
- Moon, A. (2018, June 28). Majority of Americans think social media platforms censor political views: Pew survey. *Reuters*.
- Morris, M., & Ogan, C. (1996). The Internet as mass medium. *Journal of Communication*, 46(1), 39–50. doi:10.1111/j.1460-2466.1996.tb01460.x

Ng, A. (2018, May 15). Facebook deleted 583 million fake accounts in the first three months of 2018. *Cnet News*.

Nielsen, E. (2018, Aug. 9). If We Silence Hate Speech, Will We Silence Resistance? *New York Times*.

Pal, J., Thawani, U., van der Vlugt, E., Out, W., & Chandra, P. (2018). Speaking their Mind: Populist Style and Antagonistic Messaging in the Tweets of Donald Trump, Narendra Modi, Nigel Farage, and Geert Wilders. *Computer Supported Cooperative Work*, 1–34.

Persily, N. (2017). The 2016 US Election: Can democracy survive the internet? *Journal of Democracy*, 28(2), 63–76. doi:10.1353/jod.2017.0025

Pew Research Center. (2018). *Social Media Fact Sheet, February 5, 2018*. Author.

Qualman, E. (2010). *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons.

Rahimi, B. (2011). The agonistic social media: Cyberspace in the formation of dissent and consolidation of state power in postelection Iran. *Communication Review*, 14(3), 158–178. doi:10.1080/10714421.2011.597240

Remland, M. S. (2016). *Nonverbal communication in everyday life*. SAGE Publications.

Schoop, M. E. (2017). *Independent Music and Digital Technology in the Philippines*. Taylor & Francis. doi:10.4324/9781315403267

Schram, W. E. (1954). *The process and effects of mass communication*. Academic Press.

Thomas, R. J. (2018). 20 Advocacy Journalism. *Journalism*, 19.

Weaver, K. D. (2018). Censorship in the Digital Age the World Over. In *Encyclopedia of Information Science and Technology* (4th ed., pp. 7292–7301). IGI Global. doi:10.4018/978-1-5225-2255-3.ch634

Xiang, Z., & Fesenmaier, D. R. (2016). *Analytics in Smart Tourism Design*. Springer International Pu.

ADDITIONAL READING

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *The Journal of Economic Perspectives*, 31(2), 211–236. doi:10.1257/jep.31.2.211

Bertot, J. C., Jaeger, P. T., & Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government Information Quarterly*, 29(1), 30–40. doi:10.1016/j.giq.2011.04.004

Correa, T., Hinsley, A. W., & De Zuniga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26(2), 247–253. doi:10.1016/j.chb.2009.09.003

Couldry, N. (2012). Media, society, world: Social theory and digital media practice. *Polity*.

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*, 54(3), 265–273. doi:10.1016/j.bushor.2011.01.007

- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68. doi:10.1016/j.bushor.2009.09.003
- O’Keeffe, G. S., & Clarke-Pearson, K. (2011). Clinical report—the impact of social media on children, adolescents, and families. *Pediatrics*, peds-2011.
- Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1), 59–68. doi:10.1080/15295036.2016.1266686
- Safko, L., & Brake, D. (2009). *The social media bible*. Wiley.
- Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign Affairs*, 28–41.
- Sunstein, C. R. (2018). *Republic: Divided democracy in the age of social media*. Princeton University Press.
- Wingfield, N., Isaac, M., & Benner, K. (2016). Google and Facebook take aim at fake news sites. *The New York Times*, 11, 12.

KEY TERMS AND DEFINITIONS

Advocacy Journalism: Journalism as critique and change agent (Thomas, 2018).

BBS: Bulletin board system.

Gatekeeper: Reporting information with scientific method for objective and valid results.

Mass Communication: A one-way flow from source to recipient.

Social Media: Communication portals where anyone can sign up and freely “post” messages to self-selected groups of individuals.

Breaking Fake News and Verifying Truth

11

Kazuhiko Shibuya <https://orcid.org/0000-0002-3243-4757>

ROIS, Japan

INTRODUCTION

How to distinguish between fake news or not? How to verify information whether it is true or not? The goal of this article is to review online fact-checking studies on fake news. The core of this matter is how to progress fact-checking on misleading information and vague knowledge (Viviani&Pasi,2017; Shao, et al., 2018; Vosoughi, Roy&Aral, 2018). This procedure is crucially managed by manually or some computational mechanisms. The burden of proof should be evaluated truthiness by scientific ways included objectiveness and falsifiability. It never accepts neither solipsism nor agitating manners. In addition, it should take carefully to extract the most importance from the truth whether it is manually or mechanically, and it also reconsiders those backgrounds. For sound democracy, mediated knowledge brings the facts for citizens, and it should encourage them to ponder their directions of governances in nation (DiResta,et al.,2019; Howard, et al., 2018). Using ICT, it should properly facilitate fundamental educations for them.

BACKGROUND

The world is the totality of facts, not of things (Wittgenstein, 1922). Fake news recently reminds us this renowned sentence. Wittgenstein also said that the world consists of a set of facts, and facts are determined by truth or false as a proposition style. He casted questions on world cognition, provable logics, facts and knowledge. However, to date, fake news and post-truth are too far from truth, facts, impartiality and fairness. Any assertions of fake news are never valid and those contents are almost based on malicious intentions or misunderstandings. Moreover, fake news can spread faster than other reliable information through social networks (Doer,Fouz&Friedrich,2012; Roozenbeek&Linden,2018).Especially, these mediated information which recognized as fake news can be found in the fields such as politics, medical health, socioeconomics, ethnicity, ideology, gender and other sensitive topics. Regarding populism, some governors demonstrate 'self-first' actions and they sometimes use it as political propaganda against their opponents (Bobo,2017; Lamont, Park&Ayala-Hurtado, 2017; Wang,Luo,Niemi&Li, 2017).

Lexically, Oxford dictionary defines 'post-truth' as "*relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief*". And Nielsen&Graves(2017) identified types of fake news such as *satire, poor journalism, propaganda, some advertising and false news*. Similar concepts can enumerate stereotypes (Lippmann,1922), prejudice and rumor. Moreover it definitely claims that it also lacks falsifiability (Popper, 1963) and it seems a kind of solipsism of each claimant.

DOI: 10.4018/978-1-7998-3479-3.ch101

FOCUS OF THE ARTICLE

When borrowing Marx's proverbs, a specter is haunting the Worlds -the specter of fake news. All the powers of global communities have entered into a holy alliance to exorcise this specter. At 2nd Feb. 2017, AAAS (American Association for the Advancement of Science) held in an academic symposium for discussing the fake news and post-truth. As through social media, they argued that our contemporary society has vulnerability against verification on fact and credibility of information sources. And of course, Facebook, Twitter and other social media sites have already launched eligible regulations and managements in verifying submitted contents and fact-checking.

However, there will be a stumbling block against those objectives. Renowned mathematician Hilbert believed that all statements in mathematics can be proven as true or false. But please recall the fact on Turing Machine (Turing, 1937; Copeland&Shagrir, 2019). Similarly, revealing fake news seems possible, it is not easy tasks. Whereas our constraints are within finite calculation time and efforts, fake news has been growing steadily with enormous contents and its derivations added each day. Will it be possible to verify all of them within finite time? As Turing machine proved, it is hard to demonstrate certainly truth regardless of the limits of manpower or computer. For example, if one finds the fact that it is not fake news by verifying target statements (for example, from 1 to N) within each content suspected to be certainly fake news, however it cannot be concluded that all of them are true, just because such disproof is not found in that time. In the future, by adding $N + 1$, it might turn out fake news. Then, these verifications on the bases of inductive reasoning and 'evidence of absence' are still harder for us.

The goal is the very clear. In the context of finite space, computable and manageable constraints, the core of this matter is how to progress fact-checking on those fake information and uncomfortable knowledge online (Viviani&Pasi,2017; Shao, et al, 2018; Ciamapglia, 2018). Principally, breaking fake news is to distinguish between subjective belief and objective facts. The former is trust and belief in our cognitive system and it justifies our world of view. The latter is fact and evidence, and it shall stress on scientific verification and prioritized importance rather than public preferences for acceptances. And some reasons in fake news problems such as scientific process and human mental model as belief system makes it more difficult. Can be strict matching both truths between subjective and objective perspectives consistent with each other? As extracting only important things from worthless matters in massive online sources, how to evaluate the true fact in ordered by some confident prioritizations?

In brief, any fact-checking process requires decisively at least below methodologies. If and only if empirical verified evidences and its reexamination processes are ensured, it can be regarded as canonical manners of scientific knowledge.

1. *Experimental*
2. *Detection*
3. *Estimation*
4. *Collective*
5. *Third party Authentication*

Regarding these concerns, there are reasons systematically to study fake news. That is, there are some definitive categories to consider at least.

1. Verification of Fake News
 - i. Definition on Fake news or not
 - ii. Verification Methodologies
2. Ambivalence of Social Media
 - i. Recognition on Sharp power
 - ii. Education for Ordinary Citizens

SOLUTIONS AND RECOMMENDATIONS

Experimental

Experiment can be conducted as strong and decisive methodology for clear understanding on phenomena and facts (RAND Corporation, 2017). The burden of proof should be evaluated truthiness by scientific ways included reproducibility, objectiveness and falsifiability (Fanelli, 2018; Shiffrin, Börner & Stigler, 2018). It never accepts neither solipsism nor agitating manners. According to Sober (2008), data centered experimental researches which are based on inductive evidence and facts can be categorized as Bayesian and other schools. Generally, Bayesian School specifies their studies to model observer's objective probability and these conditional facts, and then it can be properly estimated updating certainty for true nor not.

Usefully, Bayesian statistics can also detect a subjective probability of each individual when it acquires post probability information as truth or falseness (Kahneman & Tversky, 1973, 2000). They demonstrated serial consequences by psychological experiments that many of subjects unintentionally and unconsciously evaluated higher subjective probability for a member of their belonging group (in-group) than other groups (out-group). These statistical significant evidences clearly were obeyed in formalization derived from Bayesian statistics. It was a well-known breakthrough for researches in cognitive biases, and behavior science and social psychological experiments repeatedly endorsed that human being is inevitably apt to neglect their undesirable information, and they are strongly committing to believe in their supportive one (Oakes, Haslam & Turner, 1994).

Here, there was empirically observation data obtained from USA president election campaign at 2016. Before this voting day, according to a report conducted by Pew Research Center (2017), two third of Americans got some news from social media including fake news. Further, for example, fake news: "Pope Francis shocks world, endorses Donald Trump for president, releases statement" was supported by three fourth (75%) of the total of Trump supporters, whereas his opponents were only 46%. It really exhibits in-group bias for interpretation consolidated with their political ideologies. If still yet, those who eagerly accept fake news such as delusional, anti-intellectual and anti-scientific believers may be still confused.

Detection

Detecting fake news clearly relates to AI studies on developing autonomous mechanisms such as logics, recognition, representation, knowledge and inference (Russel & Norvig, 2009). Especially, AI studies generated automatic classifiers of web contents, web crawlers, autonomous bots, and agent-based services online. As standing on general computer engineering as well as security (e.g. ISO/IEC JTC1/SC27), necessary features should be referred in employing fact-check systems on fake news contents (e.g. authenticity, accountability, non-repudiation, and reliability).

Recently, Shu et al. (2017) categorized characterizations of fake news as psychological and social foundations as well as malicious accounts and echo chamber. And they proposed several methodologies for fake news detections such as *knowledge-based*, *style-based*, *stance-based* and *propagation-based one*. In this case, there is necessary to prepare ontological database, corpus and knowledgebase in advance, and natural language processing is quite needed in massive text data.

Otherwise, there are varieties of online services such as detection of bots spreading fake news (e.g. *Distil Networks*), analysis on behavior pattern of site visitors, reputation management against fake news, dark-web monitoring and block-chain based credibility of information (e.g. *Userfeeds*). And Truelove et al. (2017) proposed the ways to validate fact check against posted contents by through Twitter and other similar online services. And they found the strong correlated relationships on objective confidence for media contents when third party readers offered confirmable facts such as witness information, supporting rate by other bloggers and attaching supplements of visual information (e.g. image or thumbnail files) with geotag data (Fang, Nourbakhsh, Liu, Shah & Li, 2016). However, to date, text formatted news (e.g. bot, short blog, news headline) are automatically generative by AI based technologies (Shao, et al., 2018). Then those generated information do not provide any confidence against the evidence of fake news yet (Schafer, Evert & Heinrich, 2017).

On the other hands, facial information employs a key factor in any interactive communications in daily life and speeches in politics. And image recognition technologies have recently advantages for evidence based communication system. In this concern, Wu, et al., (2018) has produced AI based deception detecting system. They articulated that their systems have approached to high quality for detecting deception of humans' vague facial expressions and motion captured data extracted from video clips. This approach may open the possibility for deception-less interactive communications.

However new technology has already encountered new vulnerabilities against computer security and anomaly. It is the very possible to lose the evidence for occurred events provided formats by video-clip and photo files using generative adversarial network (GAN) models in AI technologies (Goodfellow, Bengio & Courville, 2016). Deep learning is specified as pattern recognition of machine learning based on massive samples. In contrast, roughly speaking, GAN is defined to recreate and modify mathematically arbitrary image files (Li, Swersky & Zemel, 2015). Thus those automated artifacts as image files can be automatically regenerated by this model to match actual events, and this technique has been worried about effective abuse for alibi and counterfeit in fake news and political forgeries against antagonists. Thus those models have possibilities to critically damage blind spots in computer security using facial recognition system for identified authentication. As smart phone and other computational devices have equipped digital cameras, and ordinary people can attach their photo clips for identification to intercommunicate with each partner, but they can make facial deceptions each other. Online witness can also make any deceptions by photo clips regenerations. Moreover, according to empirical findings in brain sciences, trust for other individual is a fundamental for our sociality, but those cognitive and neural mechanisms were not well revealed. Feldman Hall, et al (2018) found that similarity of stimulus between trusted acquaintance and morphed face image was a critical factor for trustworthy. Can our human mind break those modified deceptions?

Estimation

In computational studies for verification on fake news, it should devote robust explanations standing on mathematical and statistical relevance. These researches needs to manage and analyze online big-data and these are still higher priority. Knowledge based reasoning (Hosseinmontlagh & Papalexakis, 2018),

natural language processing and machine learning can be diversified in these days. These fields in data science and computer sciences have varieties of statistical means for the purpose of fact checking and credibility evaluation on online contents (Hastie, Tibshirani & Friedman, 2009; Liu, et al., 2017; Shao, et al., 2018). And there are needs for tracing inter-documents' links structures, contents comparisons and analyzations in ontological semantic means.

At least, to evaluate observation and choose best fitting models for purposes, it enormously offers supervised and unsupervised learning (including deep learning), Bayesian model (including Bayesian networks), classification, regression models, neural networks, clustering, tensor model, random forest, support vector machine, genetic algorithm (Goldberg, 1989) and other mathematical models. In addition, social network analysis for social media can be useful to detect opinion groups between discussing site, and it can be possible to explore various factors of inter-group conflicts or mutual cooperation. And recently, sparse modeling and pattern recognition algorithms can contribute to detect and classify fake news. The former can accurately predict tendencies based on few samples (Mairal, Bach & Ponce, 2014), and the latter is to acquire specification for identifying data.

In practical cases, there are already many tools such as artificial intelligence based machine learning, data-mining, statistical algorithm and mathematical modeling (Pramanik, et al., 2017). And some studies produced automatic detection and monitoring systems against online contents such as fake rumors, hate speeches and agitations (Ratkiewicz, et al., 2011). In contrast, it launched a kind of counter-media to impress good facets of the facts. In this result, it will expect to decrease inter-conflicts and antagonism between individuals, groups and nations (Article 19, 2015; Garimella, et al., 2017).

Collective

Crowd and collective intelligence has been recognized as one of proper ways to verify evidence (Das & Magdon-Ismael, 2012). As one of comprehensive works, Viviani & Pasi (2017) has summarized research topics on these fields. For instance, they picked up various cases: (1) detecting opinion spams in review site, (2) the detection of fake news and spam in microblogging, and (3) the credibility assessment of online health information. To evaluate reliance of information, it should include verification on fact information itself as well as information sources. As they said, users' reliance for online information can be based on 'crowd consensus'. It recalls Wikipedia editing style. For the considerable ways of proofing truth online, they categorized (1) contents analysis, (2) users' behavior analysis, (3) linkage and graph structure analysis and (4) multiple ways.

In this concern, Facebook recently tackles these matters, practically, but manual and automatic hybrid checking are still hard for verification on credibility of contents. As statistical pattern matching and machine learning on the bases of the 'reliable knowledgebase' as past similar events and facts, the fundamental of contents verification is to be coincided with mutual consistency between true or false in each information sources. However these AI based mechanisms have weak points not to easily judge the truth on statistical rare cases and individual matters (e.g. hidden personal profile). For confident validation on such things, many of crowd based experts should take participation in those processes.

In political contexts, those who have common senses among journalists to do voluntary validation on fact check are collectively verifying political utterances and speech contents by government leaders and politicians. There are already various online grass-root activities for revealing fake news and post truth of propaganda such as *Politifact*, *Hoaxy*, *Snopes.com*, *FactCheck.org*, *The Washington Post Fact Checker*, *CNN Reality Check* and others.

Similarly, as academic peer review are underlying in collective participation of external experts (Hofmann, 2018), Price&Flach (2017) argued how artificial intelligence will be progressing crucially peer review. As mentioned above, peer review on academic papers is the very similar to fake news checking. Much misconduct in academics can be perplexed in enormous efforts, but it still requires manual and hard works (Steen,Casadevall&Fang, 2013). Using AI based review systems, it actually applies to various other contents. These mechanisms can also help for identifying fake news.

Third party Authentication

Particularly, ordinary citizens are often surrounded with fake news related to health information as well as politics, and then it should be critically reexamined by further investigations. For example, Viviani&Pasi (2017) also argued fake news on medical health knowledge in social media sites. And as mentioned above, they proposed survey-based fact checking approach by trustable authorities, opinion leaders and experts. This reason requires fact checking by higher expertise and quality assurances for online contents. But it consumes much time and costs.

Thus serial efforts by third party can offer authenticity of scientific knowledge and prevent further misunderstandings of pseudoscience. Crucially, in medical health cases, there are often significant issues not to be regarded as varieties of fake news, but it definitely denotes that scientific knowledge strictly requires convincing verifications. For example, as many clinical research findings could not be often reproduced, and then NIH (National Institute of Health, U.S.A) had already recommended that the results of significant findings should be revalidated (Ioannidis, 2005; Fanelli, 2018). These reasons are that some of them are consisted of the conditions under suspicious of forgeries or ambiguous results in examinations (Steen, et al., 2013). In other cases, accepted academic papers are not always endorsed by other authorities in scientific manners. For example, it can exhibit a case to recognize an importance of risk assessment against GMO (genetic modified organization) in 2012. A published paper reported that rats who had eaten genetic modified corn tended to die faster than other control condition (Séralini, et al., 2012), but it raised controversy on scientific verification claimed from the European Food Safety Authority (EFSA). As results of this dispute, EFSA concluded original revalidation for this evidence was inadequate. Namely accepted paper about the risks of GMO in academic journals was overturned by third party authority (Houllier, 2012).

Regarding these backgrounds, fake news distributors often abuse the difficulties for verification related to medical health as well as political information. Not without saying, it is also certainly to require credibility and transparency of reviewed information in those fact-checking site organized by third party (Brandtzaeg&Følstad, 2017).

FUTURE RESEARCH DIRECTIONS

Global Affairs against Media Manipulation and Sharp Power

Due to fake news and post truth by through social media, global society might be on the brink of democracy. The era of social media had already opened the Pandora box as the cyber war (e.g. sharp power or hybrid war) as well as communication tools for ordinary citizens. To date, the keys are that social media already becomes global networking with world opinion and contagions of influence across the border of the worlds. In this sense, those collective dynamics participating by ordinary citizens sometimes cause to

lose the order of their national governance and sound conditions in democracy. As famous evidences are recalled, 'Arab Spring' and similar events had outperformed their national governing regimes (Shibuya, 2017). In democratic societies, nevertheless those revolutions will not be usually happened, it should take close look at the nature of voting behaviors at elections campaigns (Shao, et al., 2018).

Regarding these conditions, citizens has blindly tendencies to believe in such information, which apparently accepted by collective followers online. Then it still remains vulnerabilities to confuse any evidences. Those who support mediated fake news desire to believe only self-regarding 'fact' and see 'fact' as they want are likely to organize their networking with like-minded partners ('echo chamber': Sunstein, 2001). Consequently, as social media empower vigorous networking by ordinary citizens and their numerous opportunities against political and ethnic opponents (Corner, 2017; Jost, et al., 2018), those cyber politics opened the door for emerging another types of public opinion formation (Choucri, 2012; Shibuya, 2017). They neglect both objective verification and sufficient fact checks on fake news, and further they will interconnect with like-minded partners and repel their opponents (Bakshy, Messing & Adamic, 2015). Gradually, this polarization among opinion groups apparently organizes their ideology as a part of the real world. In this way, when majority who eagerly desire to support extremist's demagoguery gains total legitimacy in whole democratic society, it will be time to encompass the political power by the name of the democracy. As result of this, their nation may lose a compass for governing direction.

As 'Arab Spring' and similar events showed, the most important was that those historical events and social movements against governing regime were collapsed by human-made dynamics (Gonçalves & Nicola, 2015; Jost, et al., 2018). These were complex factors, but it was not spontaneous accident. At that time, social media became as one of political ways for opinion manipulations. Inevitably consequences, using social media, for example, third parties (i.e. terrorists and antagonists) will be possibly to spread fake news, agitate public opinions, organize opinion groups and intrude to new similar events against opponent countries by their purposes.

These are obviously to inspire practical applications (Weinberger, 2011; RAND Corporation, 2017). In USA, NSF (National Science Foundation) published their press leases at 2015 that they launched the novel interdisciplinary research grants in computational and social sciences. Otherwise DARPA (Defense Advanced Research Projects Agency, USA) also pushes their intensive programs to progress studies in both academic and military purposes such as NGS2 (Next Generation Social Science) and SocialSim (the Computational Simulation of Online Social Behavior). And they stepped further to analyze massive data obtained from online social media and to examine it by social simulation.

Thus those researches will trigger to boost applied usages based on vigorous investigations for social media influences and toxicity of fake news against target countries' public opinion formation. Reversely it also spurs to motivate further researches on media control and defenses in their own nations, and it calls more similar studies.

Educations for Citizens

Here, during the age of the cold war, Morgenthau (1978) had forecasted our future of world conditions: when ordinary citizens harden their attitudes for nationalism and exclusionism by globally intermediated communication tools (global opinion or world opinion), with enthusiasm, it may become inevitably next world war in the future. His future is our time, and political demagogues in his era can be regarded as fake news in our time. His foretelling clearly said the negative facet of social media?

Democracy has been underlying in sound judgement and opinions among sovereigns (i.e. citizens). In the past, Machiavelli ironically said “*Though the populace may be ignorant, it is capable of grasping the truth*”. However, this phrase should be ignored at once. All of citizens do not have any methods to validate mediated vague information and defend their opinions, and they are easily influenced by other influencers. Political propaganda and opinion manipulations with exposing malice neglect those process itself, and fake news and post-truth are actually vanguards to perturb our sound understanding of entire knowledge (DiResta, et al., 2019; Howard, et al., 2018).

Therefore, using ICT, it should facilitate fundamental educations for ordinary citizens (Shibuya, 2006). Some countries legitimated strict regulation acts against extreme activities of spreading fake news. For the first step, as breaking fake news and critically thinking about agitations, ordinary citizens should be cultivated by themselves for own practical behavior based on the social common standard and rationality in whole society.

CONCLUSION

This article argued online fact-checking studies on fake news. The core of this matter was how to progress fact-checking on fake information and uncomfortable knowledge. Principally, breaking fake news is to principally distinguish between subjective belief and objective facts, and it shall stress on scientific important evidence rather than public preferences for acceptances. The burden of proof should be evaluated truthiness by scientific ways included objectiveness and falsifiability. It never accepts neither solipsism nor agitating manners. In addition, it should take carefully to extract the most importance from the truth whether it is manually or mechanically, and it also ponders those backgrounds. Using ICT, it should properly facilitate fundamental educations for ordinary citizens.

REFERENCES

- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. doi:10.1126/science.aaa1160 PMID:25953820
- Bobo, L. D. (2017). Racism in Trump’s America reflections on culture, sociology, and the 2016 US presidential election. *The British Journal of Sociology*, 68(S1). Advance online publication. doi:10.1111/1468-4446.12324 PMID:29114872
- Brandtzaeg, P. B., & Følstad, A. (2017). Trust and Distrust in Online Fact-Checking Services. *Communications of the ACM*, 60(9), 65–71. doi:10.1145/3122803
- Choucri, N. (2012). *Cyberpolitics in International Relations*. The MIT Press. doi:10.7551/mit-press/7736.001.0001
- Ciamapaglia, G.L. (2018). Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*, 1, 147-153.
- Copeland, B. J., & Shagrir, O. (2019). The Church-Turing Thesis: Logical Limit or Breachable Barrier? *Communications of the ACM*, 62(1), 66-74.

Corner, J. (2017). Fake news, post-truth and media–political change. *Media Culture & Society*, 39(7), 1100–1107. doi:10.1177/0163443717726743

Das, S., & Magdon-Ismail, M. (2012). A Model for Information Growth in Collective Wisdom Process. *ACM Transactions on Knowledge Discovery from Data*, 6(2), 1–10. Advance online publication. doi:10.1145/2297456.2297458

DiResta, R. (2019). *The Tactics & Tropes of the Internet Research Agency*. <https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper.pdf>

Doer, B., Fouz, M., & Friedrich, T. (2012). Why Rumors spread so quickly in social networks. *Communications of the ACM*, 55(6), 70–75. doi:10.1145/2184319.2184338

Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2628–2631. doi:10.1073/pnas.1708272114 PMID:29531051

Fang, R., Nourbakhsh, A., Liu, X., Shah, S., & Li, Q. (2016). Witness Identification in Twitter. *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, 65–73. 10.18653/v1/W16-6210

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generation as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7), E1690–E1697. doi:10.1073/pnas.1715227115 PMID:29378964

Garimella, K., & Francisci, G. D. (2017). Reducing Controversy by Connecting Opposing Views. *WSDM '17 Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 81–90.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional.

Gonçalves, B., & Nicola, P. (Eds.). (2015). *Social Phenomena: From Data Analysis to Models*. Springer. doi:10.1007/978-3-319-14011-7

Goodfellow, I., Bengio, Y., & Courville, A. (Eds.). (2016). *Deep Learning*. The MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. doi:10.1007/978-0-387-84858-7

Hofmann, B. (2018). Fake news and alternative truths in medical research. *BMC Medical Ethics*, 19(1), 4. doi:10.1186/12910-018-0243-z PMID:29374485

Hosseinmontlagh, S., & Papalexakis, E. E. (2018). *Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles*. http://snap.stanford.edu/mis2/files/MIS2_paper_2.pdf

Houllier, F. (2012). Bring more rigour to GM Research. *Nature*, 491, 327.

Howard, P. N. (2018). *The IRA, Social Media and Political Polarization in the United States, 2012-2018*. <https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report.pdf>

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8), 696–701. doi:10.1371/journal.pmed.0020124 PMID:16060722

Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J., & Tucker, J. A. (2018). How Social Media Facilitates Political Protest: Information, Motivation, and Social Network. *Advances in Political Psychology*, 39(1), 85–118. doi:10.1111/pops.12478

Kahneman, D., & Tversky, A. (1973). On The Psychology of Prediction. *Psychological Review*, 80(4), 237–251. doi:10.1037/h0034747

Kahneman, D., & Tversky, A. (2000). *Choices, Values, and Frame*. Cambridge University Press. doi:10.1017/CBO9780511803475

Lamont, M., Park, B. Y., & Ayala-Hurtado, E. (2017). Trump's electoral speeches and his appeal to the American white working class. *The British Journal of Sociology*, 68(S1). Advance online publication. doi:10.1111/1468-4446.12315 PMID:29114866

Li, Y., Swersky, K., & Zemel, R. (2015). *Generative Moment Matching Networks*. <http://proceedings.mlr.press/v37/li15.pdf>

Lippmann, W. (1922). *Public Opinion*. Harcourt Brace and Company.

Liu, X., Nourbakhsh, A., Li, Q., Shah, S., Martin, R., & Duprey, J. (2017). *Reuters Tracer: Toward Automated News Production: Using Large Scale Social Media Data*. DOI: doi:10.1109/BigData.2017.8258082

Mairal, J., Bach, F., & Ponce, J. (2014). *Sparse Modeling for Image and Vision Processing*. Now Publication.

Morgenthau, H. J. (1978). *Politics Among Nations: The Struggle for Power and Peace*. Knopf.

Nielsen, R.K., & Graves, L. (2017). “News you don’t believe”: Audience perspective on fake news. Reuters Institute Fact Sheet.

Niklewicz, K. (2017). Weeding out fake news: An approach to social media regulation. *European View*, 16(2), 335–335. doi:10.1007/12290-017-0468-0

Oakes, P. J., Haslam, S. A., & Turner, J. C. (1994). *Stereotyping and Social Reality*. Blackwell.

Pew Research Center. (2017). *News Use Across Social Media Platforms 2017*. http://assets.pewresearch.org/wp-content/uploads/sites/13/2017/09/13163032/PJ_17.08.23_socialMediaUpdate_FINAL.pdf

Popper, K. (1963). *Conjectures and Refutations*. Routledge.

Pramanik, M. I., Lau, R. Y. K., Yue, W. T., Ye, Y., & Li, C. (2017). Big Data analytics for security and criminal investigations. *WIREs Data Mining and Knowledge Discovery*, 7(4), e1208. Advance online publication. doi:10.1002/widm.1208

Price, S., & Flach, P. A. (2017). Computational Support For Academic Peer Review: A Perspective from Artificial Intelligence. *Communications of the ACM*, 60(3), 70–79. doi:10.1145/2979672

RAND Corporation. (2017). *Priority Challenges for Social and Behavioral Research and Its Modeling*. https://www.rand.org/content/dam/rand/pubs/working_papers/WR1200/WR1206/RAND_WR1206.pdf

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. *WWW '11 Proceedings of the 20th international conference companion on World wide web*, 249–252.

Rehm, G., & Declerck, T. (Eds.). (2017). *Language Technologies for the Challenges of the Digital Age*. Springer.

Roozenbeek, J., & Linden, S. (2018). The Fake News Game: Actively Inoculating Against the Risk of Misinformation. *Journal of Risk Research*. Advance online publication. doi:10.1080/13669877.2018.1443491

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Schafer, F., Evert, S., & Heinrich, P. (2017). Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinzo Abe's Hidden Nationalist Agenda. *Big Data*, 5(4), 294–309. Advance online publication. doi:10.1089/big.2017.0049 PMID:29182493

Séralini, G. E., Clair, E., Mesnage, R., Gress, S., Defarge, N., Malatesta, M., Hennequin, D., & de Vendômois, J. S. (2012). Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. *Food and Chemical Toxicology*, 50(11), 4221–4231. doi:10.1016/j.fct.2012.08.005 PMID:22999595

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1), 4787. doi:10.1038/41467-018-06930-7 PMID:30459415

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS One*, 13(4), e0196087. doi:10.1371/journal.pone.0196087 PMID:29702657

Shibuya, K. (2006). Collaboration and Pervasiveness: Enhancing Collaborative Learning Based on Ubiquitous Computational Services. In M. Lytras & A. Naeve (Eds.), *Intelligent Learning Infrastructures for Knowledge Intensive Organizations: A semantic web perspective*. IGI. doi:10.4018/978-1-59140-503-0.ch015

Shibuya, K. (2017). Bridging between Cyber Politics and Collective Dynamics of Social Movement. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (4th ed.). IGI Global.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2632–2639. doi:10.1073/pnas.1711786114 PMID:29531095

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake news detection on social media: A Data mining perspective*. https://www.kdd.org/exploration_files/19-1-Article2.pdf

Sober, E. (2008). *Evidence and Evolution: The Logic behind the Science*. Cambridge University Press. doi:10.1017/CBO9780511806285

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why Has the Number of Scientific Retractions Increased? *PLoS One*, 8(7), e68397. doi:10.1371/journal.pone.0068397 PMID:23861902

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Article 19. 2015. 'Hate Speech' Explained: A Toolkit. (2015). [https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-\(2015-Edition\).pdf](https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-(2015-Edition).pdf)

Truelove, M., Vasardani, M., & Winter, S. (2017). Testing the event witnessing status of microbloggers from evidence in their micro-blogs. *PLoS One*, 12(12), e0189378. doi:10.1371/journal.pone.0189378 PMID:29232395

Turing, A. M. (1937). On Computable Numbers: with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(1), 230-265. 10.1112/plms2-42.1.230

Viviani, P., & Pasi, G. (2017). Credibility in social media: Opinions, news, and health information—a survey. *WIREs Data Mining and Knowledge Discovery*, 7(5), e1209. Advance online publication. doi:10.1002/widm.1209

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false online. *Science*, 359(6380), 1146–1151. doi:10.1126/science.aap9559 PMID:29590045

Wang, Y., Luo, J., Niemi, R., & Li, Y. (2017). *To Follow or Not to Follow: Analyzing the Growth Patterns of the Trumpists on Twitter*. The Workshops of the Tenth International AAAI Conference on Web and Social Media News and Public Opinion: Technical Report WS-16-18.

Weinberger, S. (2011). WEB OF WAR. *Nature*, 471(7340), 566–568. doi:10.1038/471566a PMID:21455152

Wittgenstein, L. (1922). *Tractatus Logico-philosophicus*. Harcourt, Brace.

Wu, Z., Singh, B., Davis, L. S., & Subrahmanian, V. S. (2018). *Deception Detection in Videos*. <https://arxiv.org/pdf/1712.04415.pdf>

ADDITIONAL READING

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. doi:10.1007/978-0-387-84858-7

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Viviani, P., & Pasi, G. (2017). Credibility in social media: Opinions, news, and health information—a survey. *WIREs Data Mining and Knowledge Discovery*, 7(5), e1209. Advance online publication. doi:10.1002/widm.1209

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false online. *Science*, 359(6380), 1146–1151. doi:10.1126/science.aap9559 PMID:29590045

KEY TERMS AND DEFINITIONS

Fake News: Nielsen and Graves (2017) identified types of fake news such as satire, poor journalism, propaganda, some advertising, and false news.

Sharp Power: It originally refers to a variety of cyber warfare being operated by authoritarian powers using social media and ICT. It contrasts with hard power (military forces) and soft power (socioeconomic and cultural means).

Crossfire and Violation of Human Rights in Bangladesh: A Critical Review

11

Md. Awal Hossain Mollah

University of Rajshahi, Bangladesh

INTRODUCTION

This paper depicts the violation of human rights caused by law enforcing agencies in Bangladesh. The law enforcement agencies of Bangladesh are universally appreciated for their patriotic and courageous efforts to combat terrorism inside the country. The law enforcing agencies are mainly responsible for maintaining the law and order, protection of life and property of the citizens and prevention and detection of crime to establish rule of law in a state. However, the violation of human rights by the law enforcing agencies in Bangladesh has been increasing severely for the last few years. Why the protectors become perpetrators has also explored in this paper by highlighting the pattern of violations made by the law enforcing agency and their impact upon the law and order situation in Bangladesh. The existing legal framework of governing the law enforcing agency has also been critically examined and finally pinpoint some recommendations for the eradication of the pitfalls of security forces. The paper is qualitative in nature and case study on Bangladesh.

BACKGROUND

This section focuses on the major elementary issues and background of the violation of human rights in Bangladesh. Primarily, crossfire and violation of human rights was a media term in Bangladesh, however, now widely used to express the murder of a criminal or accused in a gunfight event between members of law enforcement agencies and criminal groups. It alleged by the media, opposition political parties and civil society that law enforcing agencies are pressured to get quick results, often with unofficial guarantees that they may use any means possible to accomplish the task at hand. This occurrence is facing enormous criticisms in the home and abroad and considered as a violation of human rights.

Human Rights

The definition of human rights is a debating issue but it has been discussing as a burning issue from time immemorial. Human rights are rights inherent to all human beings, whatever our nationality, place of residence, sex, national or ethnic origin, color, religion, language, or any other status. Every single citizen is equally entitled to human rights without discrimination (United Nations, 2016). Human rights, deprivation and unequal allocation of resources have been issues for then West Pakistan (now Bangladesh) between 1947 and 1971. The people of Bangladesh struggled for 23 years to ensure and establish their rights. In 1971, Bangladesh was independent but three million people lost their lives, ten million

DOI: 10.4018/978-1-7998-3479-3.ch102

fled across the border to India and 200,000–400,000 women were raped, leading to approximately 25,000 pregnancies (Linton, 2010:194). Could anyone imagine how brutally human rights are violated in Bangladesh? Though 47 years of independence have passed, however, violation of human rights is still continuing in Bangladesh. The country struggled to overcome significant barriers to the realization of human rights: a repressive police force, poverty, corruption, unequal distribution of wealth and opportunities, and a slow legal system. The state apparatus deliberately suppressed human rights, extraterritorial actors and the dynamics of regional and global hegemony allowed limited human rights for the people (Halim, 2010).

Rapid Action Battalion

In 2004, the government of Bangladesh introduced the elite anti-crime force called the Rapid Action Battalion (RAB) following a rise in the rate of crime in the country. The RAB was created through the “The Armed Police Battalions (Amendment) Act, 2003,” and published in the Bangladesh Gazette of July 12, 2003. The primary tasks of the RAB were to maintain internal security; recover unauthorised arms, ammunition, explosives and other such articles; apprehend armed gangs of criminals; assist other law enforcement agencies in the maintenance of law and order; gather intelligence; investigate any offence concerning the government; and, any other such duties as the government requires from time to time. Seconded members from six services such as the Bangladesh Army, Bangladesh Navy and Bangladesh Air Force, the Bangladesh Police, and members of Bangladesh’s other law enforcement groups such as Border Guard Bangladesh and Bangladesh Ansars joined the RAB (Human Rights Watch, 2011). This force has been accorded with special power with special salary and benefits package. They are better equipped than the regular police force. For example, in order for them to arrive at the crime scene fast, they are given a helicopter.

Crossfire and Extra-Judicial Killing:

The term ‘crossfire’ is synonymous of the term ‘death in an encounter’ which is used by law enforcement agencies in Bangladesh. The government justifies the killings by using the term “crossfire,” which it refers to as gunfights between any alleged criminal group or terrorist and the RAB or police (Habib, 2016). The term ‘extrajudicial killing’ means killing someone without following the legal justice system, laws, and constitution (Kamruzzaman, Khan, and Das, 2016). Killings committed outside the judicial process by or with the consent of public officials, other than as necessary measures of law enforcement to protect life or as acts of armed conflict carried out in conformity with the rules of international humanitarian law is called extrajudicial killing (Rodley, 2009). Such killings are prohibited by Bangladesh’s Constitution and national laws, as well as by various international instruments that Bangladesh has ratified. However, the extrajudicial killing has been made an integral part of the law-enforcement system of Bangladesh for many years by “crossfire” or “encounter” with few other synonyms like “gunfight”, “in the line of fire” etcetera (Kamruzzaman, Khan, and Das, 2016). The overall human rights situation in Bangladesh is still at an alarming level in the first six months of 2018 due to a rapid increase of extra-judicial killings, a high number rapes and sexual assault of women and custodial deaths across the country.

FOCUS: STATUS OF HUMAN RIGHTS IN BANGLADESH

This part analyses the main text of the chapter under following sections.

Legal Framework of Human Rights in Bangladesh

Bangladesh has ratified all the major international human rights instruments such as the International Covenant on Civil and Political Rights, 1966 (ICCPR) (Article 6, 7, 9, 10) (Momtaz, 2013). International Covenant on Civil and Political Rights, 1966 ensures that no one shall be subjected to torture, or to cruel, inhuman or degrading treatment or punishment. The Convention Against Torture (CAT), Convention on the rights of the Child (CRC), and the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) (ASK, 2015). Bangladesh is a party to the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT), which stipulates that State Party shall ensure that all acts of torture are offenses under its criminal law (Momtaz, 2013). The same shall apply to an attempt to commit torture and to an act by any person which constitutes complicity or participation in torture (UNDP, 2002).

Torture by public officials date back to the period immediately following independence, and it appears that torture has become institutionalized, a practice that is perpetrated regardless of the government in power; there are only fluctuations in the nature of the torture practices and in the scale (Amnesty International, 2000). However, the constitution of Bangladesh also outlaws any form of torture (ICG, 2009). The Universal Declaration on Human Rights (UDHR) (art. 3, 5, 7, 8, 9, 10, 11) is considered as the jus cogens throughout the world from which no derivation is possible. Still 'torture' has not been defined as a crime in Bangladesh (Momtaz, 2013). However, there are a number of reservations has been made by Bangladesh which made it ineffective and denies the equal status of women. Bangladesh reserved article 14 of the Convention against Torture which states "Each State Party shall ensure in its legal system that the victim of an act of torture obtains redress and has an enforceable right to fair and adequate compensation, including the means for as full rehabilitation as possible (Momtaz, 2013). In the event of the death of the victim as a result of an act of torture, his/her dependents shall be entitled to compensation. Therefore, the implications for this reserve are to deny such compensation or rehabilitation and also invalid the constitutional mandate under Article 35 (5) of the People's Republic of Bangladesh (United Nations, 2016). The International Covenant on Civil and Political Rights (ICCPR) which was ratified by Bangladesh in 5th September 2000 ensures that every human being has the inherent right to life and that this right should be protected by law and no one shall be arbitrarily deprived of his life (Momtaz, 2013). The Bangladesh Constitution also provides provisions to ensure human rights.

The Constitution of Bangladesh has been grunted 18 fundamental rights by Articles 27 to 44. Articles 31 and 32 guarantee the right to life and personal liberty and it is applicable to both citizens of Bangladesh and foreign nationals (Habib, 2015). Article 33, provides that the arrested person must be produced before a magistrate within 24 hours excluding the time of journey from the place of arrest to the court of the magistrate.

The HCD of Bangladesh held that the detention of prisoners in Jail after they have served out the sentences is a violation of their fundamental rights guaranteed in the Constitution, especially the right to life (Law Help Bangladesh, 2016; Momtaz, 2013). Hence, through this judicial activism, there is scope for a post-trial concession to enjoy by a prisoner who has spent in custody for a considerable time (Momtaz, 2013).

The constitutional guarantee of the right to life – the right not to be deprived of life is given in accordance with the law and cannot be suspended or derogated from under any circumstances (Habib, 2016). However, in earlier years, state security forces reportedly continued to be responsible for extra-judicial killings in the form of cross-fire deaths, death in encounter/encounter killings and custodial killings Momtaz, 2013). These incidents have become so frequent that a new term has been coined in the country: death by crossfire. Currently, the most significant human rights problems are enforced disappearances. Human rights groups allege the involvement of the country's security forces (United States Department of State, 2012).

LAW ENFORCING AGENCIES IN BANGLADESH

In Bangladesh, as a law enforcing agency we have the 'police force'. The year 2004, three special units were formed to help the regular police. These are:

- Rapid Action Battalion
- Cheetah; and
- Cobra.

For the better understanding of our discussion, the legal basis of our law enforcing agencies is given below:

Bangladesh Police Force

Queen Victoria has ascended the throne of this Indian Subcontinent in 1858 and during her reign the Indian Police Act, 1861 was enacted (Momtaz, 2013). And the existing police force as an organized department was established during the British rule in 1861 through the promulgation of this Police Act (Act V of 1861). Although there have been a few amendments while it was renamed the Police Regulation of Bengal or PRB, the Act has virtually remained the same in spirit and still in force after almost 150 years (Momtaz, 2013).

Rapid Action Battalion

The Rapid Action Battalions (hereinafter 'RAB'), the elite force was created through 'the Armed Police Battalions (Amendment) Act, 2003 and published in the Bangladesh Gazette of July 12, 2003 (Habib, 2016). The RAB is a composite force, manpower wise, because section 3(5) of the law states: 'Notwithstanding anything contained in this Ordinance or in any other law for the time being in force, armed personnel and officers in Rapid Action Battalion may, as and when necessary, be appointed on secondment or deputation as the case may be, from among the persons who are in the service of the Republic, including any discipline force' (Kamruzzaman, Khan, and Das, 2016). It is a hybrid force consisting of the police and the army. Momtaz (2013) mentioned that The Amended Armed Police Battalion ordinance has given birth to:

- Armed Police Battalions,
- Rapid Action Battalions.

- internal security duties,
 - a. recovery of unauthorized arms, ammunition, explosives and such other articles as the government may, from time to time direct,
 - b. the apprehension of armed gangs of criminals,
 - c. assisting the regular law enforcing agencies including the Police for maintaining law and order, and such other duties as the government may, from time to time, assign.

In addition, Cheetah and Cobra were formed as special units of the detective branch. They also enjoy a lot of impunity in arresting people (Momtaz, 2013). We have, also Border Guard of Bangladesh (BGB), Industrial police, Highway police, CID, SB, 11 Training institutes, 6 Metropolitan Police, 7 Range, Railway Range, the Special Women Police Contingent (SWPC) established in 2008, and Police Hospital.

THE LAWS IN QUESTION AND HOW THEY ARE MISUSED?

Section 54 of the Code of Criminal Procedure

Under section 54 of the Code of Cr. Pc. 1898, individuals may be arrested under suspicion of criminal activity without any order from a magistrate or a warrant (Momtaz, 2013).

4.2 Section 100/86 of the Dhaka Metropolitan Police Ordinance Momtaz (2013) mentioned that Section 86 of the Dhaka Metropolitan Police (DMP) Ordinance provides:

If any person is found in the following situations during the period between sunset to sunrise:

- Equipped with dangerous machinery without satisfactory reasons, or
- Covered face or in disguise without satisfactory reason, or
- Present in any residence or any other building or any boat or water vessel or any other transport without satisfactory reason, or
- Sleeping or wandering on any street or in any place without satisfactory reasons, or
- Possessing the tools of entering any house forcefully without satisfactory reason, then the person will be liable for one-year imprisonment or fine of taka two thousand or both.

Section 100 of the DMP ordinance provides: in the presence of any police officer or in his attention if any person commits an offense under this Ordinance or any other law in force, that police officer can arrest him without a warrant.

The Special Powers Act 1974

The Special Powers Act 1974 (hereinafter SPA) is deemed by human rights activists and other members of civil society as a repressive law (Kamruzzaman, Khan, and Das, 2016). The use and abuse of this Act in the name of protecting the security of the state has resulted in a steady pattern of human rights violations. The Act provides 'Special Measures' for the prevention of prejudicial activities, for a speedy trial and punishment for grave offenses (Momtaz, 2013). Momtaz (2013) mentioned that under prejudicial acts, the law includes acts that are prejudicial to:

- the sovereign and/or state; to the maintenance of friendly relations between other states and Bangladesh;
- to the security of Bangladesh;
- to the safety or maintenance of public order;
- to the peaceful co-existence among different classes and section of the community;
- to the maintenance of law and order;
- to the security of any section of the public;
- to the economic and financial interests of the state;

For aforesaid grounds police can arrest any person at any time.

CROSSFIRE AND VIOLATION OF HUMAN RIGHTS IN BANGLADESH

The term “death in an encounter” is used in other countries which mean “crossfire” by law enforcement agencies in Bangladesh. The evil connotation associated with the word demonstrates the utter powerlessness of the people facing extrajudicial killings that are taking place in Bangladesh (Habib, 2015). The overall human rights situation in Bangladesh remained at an alarming level in the first six months of 2018 due to a sharp rise in cases of extra-judicial killings and custodial deaths, and a high number of cases of rapes and sexual assault of women across the country. A report published by the Daily Star disclosed that a total of 275 people have been killed during incidents of ‘gunfight’ and ‘crossfire’ with law enforcers, in custody of law enforcement agencies, or being shot by law enforcers, while 427 women were raped across the country between January 1 and June 30, 2018 (Daily Star, June 2018). The human rights monitoring report of Odhikar published in December 2018 mentioned that 456 people have killed extra-judicially by law enforcement agencies and 83 others were forcibly disappeared across the country by 11 months. Out of the 449 people were killed by crossfire (Odhikar, 2018). Besides, violation of human rights happened in various forms such as- death in Jail(64), political killing (79) and injured (3826), rape (619), killing by BSF (10), acid violence 26 and so may (Odhikar, 2018). The following table-1 represents the statistics of crossfire happened from 2001- to 2018 by various law enforcing agencies in

Table 1. Statistics of Crossfire 2001-2018

Crossfire/Gunfight from 2001-2018																	
Year (s)	RAB	Police	RAB- Police	Joint Force	Cheetah- Cobra	Army	BGB (Former BDR)	Police- BGB	RAB- Coast Guard	RAB-Police- Coast Guard	Coast Guard	DB Police	Forest Guard	Police- Coast Guard	Ansar	BGB- RAB	Grand Total
2018 (Jan-May)	64	134					1				2	15					216
2017	32	104				1						2					139
2016	51	92		1				1	4			2					151
2015	48	97		1			2										148
2014	23	83		8							3					2	119
2013	27	36	0	2	0	0	0										65
2012	40	7	2						4								53
2011	42	15	4						4								65
2010	65	21	9						3	3							101
2009	38	63	25			3											129
2008	65	50	15	1			1				4						136
2007	90	35	3	1							1						130
2006	181	96	1								7	4	1				290
2005	107	212			4							2		15			340
2004	63	91		7		1	1	3									166
2003	0	26		4		1						1			1		33
2002	0	5															5
2001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	936	1167	59	25	4	6	5	4	15	3	17	26	1	15	1	2	2286

Source: Odhikar Report June 2018. Statistics of Crossfire 2001-2018, http://odhikar.org/wp-content/uploads/2018/06/Statistics_Crossfire-Gunfight_2001-May-2018.pdf

Bangladesh. The table indicates that a total number of crossfire incidents happened 2286 by 18 years in which 1167 by police and 936 by RAB. It has been displayed by the table that in 2001 the violation of human rights by crossfire was zero but it reached 216 in 2017. The highest number of crossfire happened in 2005 which was 340. Though the number fluctuates but the crossfire incidents of the first 6 months of 2018 were 216 and at the end of December 2018 it reached number 449 which indicates the trend of crossfire and extrajudicial killing is increasing (Odhikar, 2018).

Bangladesh Prime Minister vows to continue anti-drug operations and over 120 people have been killed by crossfire and at least 10,000 arrested since May 14 to May 2018 by the Bangladeshi police (Arab News, 2018). This is a grave violation of human rights and a serious threat to human rights. Therefore, it has been proved that the arbitrary and abuse of power exercise by law enforcing agencies is increasing recklessly. This is the sign of lack of accountability, rule of law, good governance and deteriorating law and order and violation of civil rights in Bangladesh.

Though Bangladesh has ratified major legal instruments of International Humanitarian laws and the Constitution of Bangladesh enumerate the essential rules for protecting human rights, however, human rights always violated in various forms. Some are analyzed below. Annexes-1,2,3 and 4 shows the statistics of various forms of human rights violations till December 2018 in Bangladesh.

Death in Custody

Custodial death becomes a common happening for quite a long time. The word ‘custody’ implies guardianship and protective care. No civilized law permits custodial torture that springs out of a vicious desire to cause suffering when there is no option of any reprisal. Jails are overcrowded beyond capacity, as a result of arrest operations carried out to suppress opposition political parties and dissenting voices. The actual number of inmates is about three times more than the prisons’ capacity. The total capacity of prisons across the country is 36,614, but there were 92,404 inmates as of 30 November 2018. Many inmates allegedly die in jail due to the lack of proper treatment facilities and negligence by prison authorities. In November 2018, seven persons died in jail due to alleged ‘illnesses’. In *Jatiyo Mahila Ainjibi Samity vs. Bangladesh and Others* (criminal), the High Court held that detention in safe custody against the will of a detained person was illegal.

Many police officers believe that information or confessions cannot be extracted without physical threat or torture. That is why, although torture is prohibited under the law, its practice is rampant particularly under the ‘remand’. The form of torture includes beating on the soles of the feet, hanging, and suspension by the arms while they are tied behind the back, suspension upside down, beating, electric shocks, and psychological forms of torture including humiliation, threat, and insult etc. The Bangladesh Penal Code, 1860 has provided definitions and penalties for some offenses which very narrowly cover the area of torture as follows: offence affecting life (Section 299-311); hurt (Section 319-338A); to wrongful restraint and wrongful confinement (Section 339-348); offences relating to criminal force and assault (Section 339-358); offences relating to rape (Section 375, 376).

When such offenses are caused by the members of the police force, complains of torture can be brought against them under these provisions.

Very often it seems that torture is committed on a person on the orders of ‘higher-ups’ which allows police to act with impunity. The lack of disciplinary action against law enforcement personnel remains one of the main attributing factors of torture. Similarly, they should be instructed that any other forms of sexual abuse may constitute torture or cruel, inhuman or degrading treatment and that offender will be brought to justice’.

While addressing the protection against torture and degrading punishment under Article: 35(5) the HCD in *Alhaj Yousuf Ali vs. The State* asserted against the police torture and spoke of duty on part of the police not to exercise their power of arrest capriciously and fancifully. Torture for extracting any kind of confession would neither be right nor just nor fair’.

Arbitrary Arrest and Detention

Another form of human rights violation is arbitrary arrest and detention. In *BLAST and others vs. Bangladesh and others* [‘Section 54 Guidelines Case’, or ‘Rubel Killing Case’ or ‘Guidelines on Arrest and Remand Case’], the High Court recommended the some directives in exercising power under section 54 and 167 of the Cr Pc, 1898 and section 33 of the Special Powers Act 1974. Few can be mentioned here.

1. No Police officer shall arrest anyone under Section 54 for the purpose of detention under Section 3 of the Special Powers Act, 1974.
2. A police officer shall disclose his/her identity and show his/her ID Card on demand to the person arrested or those present at the time of arrest.
3. The police officer arrested under Section 54, or the Investigating Officer taking a person to custody or the jailor must inform the nearest Magistrate about the death of any person in custody in compliance with these recommendations
4. The Magistrate shall inquire into the death of any person in police custody or jail as per the recommendations. Later on in *Saifuzzaman vs.State* addressing Article 33 and Article 35(5), more guidelines have been issued regarding the police behavior while arresting a person as follows:
5. The police officer making the arrest of any person shall prepare a memorandum of arrest immediately after the arrest and such police officer shall obtain the signature of the arrestee with the date and time of arrest in the memorandum.
6. Time and place of arrest and the place of custody of the nearest relative or to a friend of the arrestee. An entry as to the ground of arrest and name of the informer must be made.

However, arbitrary arrest and detention are still continuing all over the country. On 8 November 2018, police arrested some people in the name of recovering drugs in Matikumra Village under Jhikargachha Upazila in Jessore District. Later the police inflicted torture on two persons named Faruk Hossain and Ashraful Islam and shot them in their legs. Faruk and Ashraful were admitted to the National Institute of Traumatology and Orthopaedic Rehabilitation in Dhaka. Dr. Ashim Ghosh said that the left legs of both Faruk and Ashraful had to be amputated. See pictures (annex-3).

Enforced Disappearance

The enforced disappearance of persons has been practiced since the mid 20th century. Enforced disappearances constitute an extreme violation of human rights, infringing many fundamental rights. Furthermore, not only is the person affected who becomes the direct victim, but also his or her family, colleagues, and friends as well as society as a whole. Most of the disappeared are persons who are critical of their governments. According to Odhikar Report (December 2018), in November 2018, 12 persons were allegedly disappeared after being picked up by members of law enforcement agencies. Among them, three were showed arrested after a few days of disappearance and the whereabouts of nine persons remain unknown. A highly visible and worrying number of enforced disappearances have been committed during the current

government reign. Opposition parties have expressed their concern and fear that there are possibilities that leaders and activists of the opposition will be disappeared in the lead up to the upcoming national elections. Many leaders and activists of the opposition parties, particularly the BNP, became victims of enforced disappearance before and after the controversial 10th Parliamentary elections in 2014. Of them, many have still not returned. The International Federation for Human Rights, Asian Federation against Involuntary Disappearance and its local member Odhikar has already urged the Bangladesh government to ratify the International Convention for the Protection of All Persons from Enforced Disappearances.

Police and Children

Children are also a victim of police's cruel behavior. In times of political unrest, children are used for throwing explosive substance and stones. That is why police always have a tendency to arrest children under S/54 and S/86 Dhaka Metropolitan Police Ordinance. In the Code of Conduct for Dhaka Metropolitan Police as to the responsibilities of the guard with the children, it is laid down that, 'Children will have to be behaved with soft words and in affectionate manners.'

As to the responsibilities of the officer-in-charge, the said Code cited that, 'in case of children or adolescents, officer-in-charge will behave with them with a vascular affection.' Many of the police stations in Dhaka city do not have separate custody room for children or female detainees. They are put in a police cell with adults and common criminals, which is, in total contravention with the Jail Code and the Children's Act' 1974. Very often after the arrest, what police doing is that they have not recorded the actual age of the children or mentioned extended age.

Under the ICCPR, children are generally considered to be individual' that enjoy the full range of rights under the covenant. Article 10(3) requires the separation of 'juvenile offenders'. But all these provisions are not ensured in Bangladesh.

Violence Against Journalists

At present, a number of journalists faced torture, threats or harassments from law enforcement agencies, (gangsters), government officials, or the student wings of various major political parties as noted in the table-4(Annex-4). Journalists are now very proactive to disclose everything before the public. That's why the government and political elites are suppressed them under recently crated laws in the name of yellow journalism by law enforcing agencies.

PROBLEMS, SOLUTIONS AND RECOMMENDATIONS

Why such violations are rampant in Bangladesh security forces? Numerous causes are responsible for such violations discussed with solution below:

Combating Corruption

Our police department has been charged against involving in corruption. Taking bribes, harassing innocent people, links with notorious gangsters, extortions, misusing power for personal gains are some of the common forms of crimes police are accused of (Momtaz, 2013). According to the report of the Transparency International Bangladesh (TIB), Corruption Perception Index (CPI), the law enforcement

agency is found the most corrupt sector in 2017 (TIB, 22 February 2018). Therefore, law enforcing agencies should be under accountability to combat corruption. Anti corruption commission may play a vital role in this case.

Salary Structure of Law Enforcing Agencies Should be Rational

The salary structure of the police department was comparatively so poor that it was simply impossible to lead an honest life. As a result, law enforcing agencies involved with a bribe and contract money and does any kinds of criminal activities. For example- the commanding officer of Rab-11 in Narayanganj Lt. Col. Tareque Sayeed Mohammad and the then company commanders Maj Arif Hossain and Lt. Commander Masud Rana of the RAB contract with Tareque is the son-in-law of Disaster Management and Relief Minister Mofazzal Hossain Chowdhury Maya. They abducted Seven people, including NCC panel mayor Nazrul Islam and senior lawyer Chandan Sarker, from the Dhaka-Narayanganj link road on April 27, 2014. Later, their bodies were found floating in the Shitalakkhya river (The Daily Star, 23 August 2017). Though recent pay scale and other facilities provided by the government (Hasina, 2014-onward) is very handsome (Basic 8250-78000) determined by National Pay Scale 2015 for the lowest to the highest level of Police officers. However, bribe or contract money still prevailing law enforcing agencies and act as a cause of crossfire and extrajudicial killing. Therefore, the salary structure should be rational based on market price of daily life.

Need Human Rights Training

One of the pitfalls lies with the existing training facility of the police. The amount of training of police officers in police stations, compared to training received at headquarters, is very low (Momtaz, 2013). A survey on the training of 100 constables of Dhaka Metropolitan Police, selected at random, who had completed 20 years of service, revealed that 96 received only basic training for 6 or 3 months (Momtaz, 2013). The basic cause is heavily biased towards physical training. Of 1,329 classes during a 6-month basic course, only 530 were devoted to academic training, and this was mostly on laws and procedures (1996). "In Bangladesh, there is no such institute for police where international standard education and training for improving overall human rights knowledge and capability of the police officers and training focusing on human security issues of the poor, woman and children can be imparted (Momtaz, 2013). As a result, of absence and ignorance, protecting and promoting the universal norms of human rights, police itself, become the perpetrators and human rights, violators.

Need Adequate and Healthy Accommodation

Accommodation is another big problem according to PRB, only 20% of the constables and sepia are allowed to live with their families (Momtaz, 2013). Others have to live in the thanas or barracks. A recent survey of accommodation by police headquarters revealed that 14,153 constables/ naiks/ head constables did not even have barrack accommodation (Momtaz, 2013). There were 32,454 cases of shortages of family accommodation for assistant sub-inspectors / head constables /naiks / constables; 4,603 for sub-inspectors and equivalents; 1,285 for senior assistant police superintendents/ assistant police superintendents/ inspectors and 98 for superintendents of police/ additional/ superintendents of police (Momtaz, 2013). Inside the dingy quarters, policemen sleep on the floor in very little space sharing

pillow and blankets. Then it is really a question of how they are to be conscious of human rights when they themselves are receiving inhuman treatment.

Recruitment Procedure Should Be Faire and Transparent

‘There is also a huge flaw in the recruitment procedure, which has far-reaching effects on the entire department. In this system, the field level officers do not have the opportunity to be promoted to the highest office in the department (Momtaz, 2013). Most of the higher level posts are exclusively set aside for officers who are appointed through the BCS Examination and start as an ASP (Momtaz, 2013). This is a strange procedure practiced in our country. This discriminating practice gives rise to frustration among the field level officers. Another important thing is that, in selecting a candidate political affiliation, good connection or how far the bribe becomes the deciding factor. By recruiting a police officer through pay off, the first thing he learns is corruption (Momtaz, 2013). So, recruitment process should be faire and transparent.

Legal Enforcement is Essential for law Enforcing Agency

There is no separate authority or department to investigate the human rights violations related to crimes committed by the police. As a result, it is the police which investigate allegations of human rights violations committed by their departmental fellow colleagues by which in most of the cases police get relief from the allegations or are awarded comparatively light punishment (Momtaz, 2013). The Armed Police Battalions Ordinance, 1979 laid down that no suit, prosecution or other legal proceedings shall be against any member of the Force, for anything which is done or intended to be done in good faith under this Ordinance (Habib, 2016). ‘That is why there are overall despair and dissatisfaction against the police so much that people are usually reluctant to report crimes in Thana as they deem sure that they will not get relief or fair treatment. No one is above the law, so police should be under control the law and judiciary.

Strong Political will

Members of the law enforcement agencies are enjoying impunity as the government is using them to suppress its political opponents (Momtaz, 2013). As a result of such impunity, allegations of harassment, torturing people, taking bribes, shooting in the legs, attacks and unlawful detention of the opposition and dissenters and extortion, were found against members of law enforcement agencies (Momtaz, 2013). The Torture and Custodial Death (Prevention) Act, 2013 was passed after relentless demands from human rights defenders. However, a vast majority of the torture victims and family members are not able to file any cases under this Act due to fear of reprisals; and those cases that have been filed are yet to see light in the court (Momtaz, 2013). Therefore, attitude of ruling political parties should be changed and at the same time strong political will is essential to punish the offenders.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

There is no doubt that Bangladesh police has been contributing to the law and order situation in Bangladesh. However, some controversial activities are ruining the image of this force. There are hundreds of allegations against the elite force officials for being involved in extortion, abduction, extrajudicial killing

by crossfire, rape and dealing in drugs and weapons. Some of the members of the RAB use organized blackmailing rackets to trap innocent people and finally extract money from the victims through various excuses. There are allegations that the law enforcing agencies have been politically used by successive governments. As discussed above those constitutional provisions says arrestees detained should be produced before the court within 24 hours after arrest. In practice, however, during the detention arrestees are not allowed to meet with anyone.

Governments and politicians have used the Bangladesh Police to destroy political enemies (ICG, 2009). It is most important that democracy is reinstated in Bangladesh but the free, fair, credible and participatory election is absent. Besides, active and participatory parliament is essential first which is lacked in Bangladesh. Accountability of government is important to establish the rule of law and good governance but all the mechanisms of accountability are under the government. Press freedom should be ensured and sufficient space should be granted for growing civil society but which is still under government control. Laws have greater permanence than commands which are also absent. The supremacy of laws ensures no person can claim to be above law but everything is strongly guided by the government. There cannot be a more undesirable situation where security forces, created for the purpose of providing safety to the people, would come to be considered by many as a source of their insecurity. Prescriptions to overcome these pitfalls are given below:

- First, we need to ensure free, fair, credible and participatory election.
- Human rights should be given the highest priority and this information has to be passed on to every single citizen.
- Parliament should be viable, vibrant, active and participatory.
- Mechanisms of accountability of law enforcing agencies and the executive branch should be more effective and free from government control.
- The culture of impunity must be seized from the existing system.
- It is to ensure that no one is above the law.
- Crossfire is not the solution to any kinds of criminal activities.
- The consciousness of people and moral education is essential for every sphere of our life.
- Lack of trust and conflicting attitude from and among political parties should be driven out.

REFERENCES

Arab News. (2018). *Bangladesh Pm Vows to Continue Anti-Drug Operations*. <https://www.arabnews.com/node/1312726/world>

ASK. (2015). *Death By Law Enforcement Agencies, 2013. January 11, 2014 - Death by Law Enforcement*. ASK.

Choudhury Salah Uddin, S. (2008). Dying while in custody: Bangladesh's Rapid Action Battalion Licensed to kill! *Canadafreepress*. <https://canadafreepress.com/article/bangladeshs-rapid-action-battalion-licensed-to-kill>

Habib, A. Z. M. A. (2015). Extrajudicial Killing in Bangladesh: A Murder of Human Rights. *International Journal of Sustainable Development*, 08(02), 67–80.

Halim, M. A. (2010). *Constitution, Constitutional Law and Politics: Bangladesh Perspective*, 2nd. CCB Foundation.

Helal, M. F., & Hafizur, R. S. (2014). Identifying The Problems Of Bangladesh Police & Rapid Action Battalion: Recommending Way. *DIU Journal of Humanities and Social Science*, 2(July), 13–34.

Human Rights Watch. (2011). *Crossfire - Continued Human Rights Abuses by Bangladesh's Rapid Action Battalion*. Available at: <https://www.hrw.org/node/98591/section/2>

ICG. (2009). *International Crisis Group, Bangladesh: Getting Police Reform on Track*. Available at <https://www.refworld.org/docid/4b22758b2.html>

Inter Press Service. (2010). *Bangladesh: No End in Sight for Extrajudicial Killings*. <http://ipsnews.net/print.asp?idnews=50053>

International Crisis Group (ICG). (2009). *Bangladesh: Getting Police Reform on Track*. Available at: <https://www.refworld.org/docid/4b22758b2.html>

Kamruzzaman, M., Khan, M. B. U., & Das, S. K. (2016). Extrajudicial Killings: The Violation of Human Rights in Bangladesh. *International Journal of Environmental Planning and Management*, 2(4), 20–28.

Law Help Bangladesh. (2016). *Criminal laws and the constitution of Bangladesh*. Law Help Bangladesh.

Linton, S. (2010). Completing the Circle: Accountability for the Crimes of the 1971 Bangladesh War of Liberation. *Criminal Law Forum*, 21(2), 191–311. doi:10.1007/10609-010-9119-8

Nora, A. (2016). *Extra Judicial Killing and Violation of Human Rights in Bangladesh*. <https://sufibor-shan.wordpress.com/2016/01/03/extra-judicial-killing-and-violation-of-human-rights-in-bangladesh/>

Odhikar. (2010). *Human Rights Report 2009: Odhikar Report on Bangladesh*. http://www.odhikar.org/documents/2009/English_report/HRR_2009.pdf

Odhikar. (2013). *Human rights report 2013*. Dhaka: Odhikar. [odhikar. org/programs/aj/police/papers/gpj/police_accountability_in_india.pdf](http://odhikar.org/programs/aj/police/papers/gpj/police_accountability_in_india.pdf)

Odhikar. (2018). *Odhikar's Human Rights Monitoring Report on Bangladesh January to November 2018*. Retrieved from: http://odhikar.org/wp-content/uploads/2018/12/human-rights-monitoring-report-November-2018_Eng.pdf

Police, B. (2016). *Crime statistics comparative*. <http://123.49.38.132/Crime-Statistics-comparative.php?id=208>

Rahman Redwanur, M. (2006). Human rights, health and the state in Bangladesh. *BMC International Health and Human Rights*, 6(1), 4. doi:10.1186/1472-698X-6-4 PMID:16611360

Report, O. (2018). *Statistics of Crossfire 2001-2018*. http://odhikar.org/wp-content/uploads/2018/06/Statistics_Crossfire-Gunfight_2001-May-2018.pdf

Rodley, N. S. (2009). *The Treatment of Prisoners under International Law*. 252. Oxford University Press.

Suraya, M. (2013). Human Rights Violations in Bangladesh: A Study of the Violations by the Law Enforcing Agencies. *Mediterranean Journal of Social Sciences*, 4(13), 101–119.

The Daily Star. (2017). *Death for 15, life term for 11*. <https://www.thedailystar.net/frontpage/narayanganj-7-murder-verdict-high-court-bangladesh-death-reference-appeal-death-15-life-term-11-1452607>

TIB. (2018). *Bangladesh in Corruption Ranking: It Could Have Been Better*. <https://www.ti-bangladesh.org/beta3/index.php/en/articles-op-eds/5507-bangladesh-in-corruption-ranking-it-could-have-been-better>

Transparency International Bangladesh. (2012). *National Household Survey*. http://www.ti-bangladesh.org/research/Executive_Summary_23122010_FINAL.pdf

UNDP. (2002). *Violence and Repression against Women and Children: An Unending Agenda, Human Security in Bangladesh, in search of Justice and Dignity*. UNDP.

United Nations. (2016). *What are human rights*. Geneva: United nations Human Rights Office of High Commissioner.

United States Department of State. (2012). *Country Reports on Human Rights Practices for 2012 United States Department of State, Bureau of Democracy*. Human Rights and Labor.

ADDITIONAL READING

Habib, A. Z. M. A. (2015). Extrajudicial Killing in Bangladesh: A Murder of Human Rights. *International Journal of Sustainable Development*, 08(02), 67–80.

Helal, M. F., & Hafizur, R. S. (2014). Identifying The Problems Of Bangladesh Police & Rapid Action Battalion: Recommending Way. *DIU Journal of Humanities and Social Science*, 2(July), 13–34.

Rahman Redwanur, M. (2006). Human rights, health and the state in Bangladesh. *BMC International Health and Human Rights*, 6(1), 4. doi:10.1186/1472-698X-6-4 PMID:16611360

Rodley, N. S. (2009). *The Treatment of Prisoners under International Law*. 252. Oxford University Press.

Suraya, M. (2013). Human Rights Violations in Bangladesh: A Study of the Violations by the Law Enforcing Agencies. *Mediterranean Journal of Social Sciences*, 4(13), 101–119.

KEY TERMS AND DEFINITIONS

Crossfire: The term ‘crossfire’ is synonymous of the term ‘death in an encounter’ which is used by law enforcement agencies in Bangladesh.

Death in Custody: Death in custody is a death of a person in the jail, in prison or custody of the police or other government law enforcing authorities.

Enforced Disappearance: In international human rights law, a forced disappearance (or enforced disappearance) occurs when a person is secretly abducted or imprisoned by a state or political organization, or by a third party with the authorization, support, or acquiescence of a state or political organization, followed by a refusal to acknowledge the person’s fate and whereabouts, with the intent of placing the victim outside the protection of the law.

Extrajudicial Killing: Extrajudicial killing is the killing of a person by governmental authorities, law enforcing agencies or individuals without the sanction of any judicial proceeding or legal process.

Human Rights: Human rights are rights inherent to all human beings, whatever our nationality, place of residence, sex, national or ethnic origin, color, religion, language, or any other status. Every single citizen is equally entitled to human rights without discrimination (United Nations, 2016).

Law Enforcing Agencies: In Bangladesh law enforcing agencies includes, Police, Army, Border Guard of Bangladesh (BGB), Industrial police, Highway police, CID, SB, 11 Training institutes, 6 Metropolitan Police, 7 Range, Railway Range, and the Special Women Police Contingent (SWPC).

RAB: Rapid Action Battalion (RAB) is an elite force created through the The Armed Police Battalions (Amendment) Act, 2003 to control crime in Bangladesh.

APPENDIX 1

Table 2. Statistics of Human Rights Violations: January-November 2018

Type of Human Rights Violation			Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.
Extrajudicial Killings	Crossfire		18	06	17	28	149	50	69	24	35	19	34
	Shot to death		1	1	0	0	0	0	0	0	0	0	0
	Torture to death		0	0	1	1	2	0	0	0	1	0	0
Total			19	7	18	29	151	50	69	24	36	19	34
Enforced Disappearances			6	1	5	2	1	3	5	5	30	13	12
Death in Jail			6	5	9	7	8	5	7	4	2	4	7
Human Rights Violation by Indian BSF	Bangladeshis Killed		2	1	0	0	0	0	1	0	1	3	2
	Bangladeshis injured		3	5	1	2	0	1	0	1	7	0	0
	Bangladeshis Abducted		2	0	0	3	4	0	0	0	1	2	1
	Total		7	6	1	5	4	1	1	1	9	5	3
Attack on Journalists	Injured		12	6	1	2	3	1	3	12	1	3	1
	Assaulted		1	3	3	0	0	0	0	10	1	0	0
	Threatened		2	1	3	0	1	1	0	1	0	0	0
	Total		15	10	7	2	4	2	3	23	2	3	1
Political Violence	Killed		9	5	9	11	13	2	3	2	4	10	11
	Injured		619	424	335	428	297	153	216	252	261	380	461
Dowry related violence against Women			12	16	15	21	12	06	10	14	16	07	06
Rape			46	78	67	69	78	48	59	56	54	53	31
Sexual harassment of Females			15	14	25	24	19	06	11	08	16	09	05
Acid violence			2	1	3	4	2	0	5	6	1	1	1
Public lynching			5	6	8	2	5	2	4	3	6	4	3
Situation of Workers	RMG workers	Killed	0	0	1	0	1	0	0	0	0	0	0
		Injured	20	0	40	0	35	27	10	0	1	67	20
	Workers in other sectors	Killed	9	11	7	8	18	7	4	6	5	3	9
		Injured	8	4	0	3	4	3	9	0	6	20	8

Source: Odhikar's Human Rights Monitoring Report on Bangladesh January to November 2018. Retrieved from: http://odhikar.org/wp-content/uploads/2018/12/human-rights-monitoring-report-November-2018_Eng.pdf

The above-mentioned data disclosed the overall picture violation of human rights caused by various ways including by law enforcing agencies by 11 months in Bangladesh. Though Bangladesh now improving the economic conditions progressively day by day the situation of rule of law, human rights, and democratic practice are deteriorating by an alarming rate.

Politics and Ethics and a National Framework to Combat Corruption in Zimbabwe

Tawanda Zinyama

University of Zimbabwe, Zimbabwe

INTRODUCTION

One of the issues which have pre-occupied the minds of researchers in Zimbabwe has always been the issue of governance and corruption. This is because governance defines the nature and outcomes of the relationship between the governors and the governed. If it follows the dictates of good governance, it brings joy to the governed, but if it is inspired by bad politics, it leads to the suffering of the people. Evidence of declining public trust has been seen in negative perceptions of the government as an institution prone to corruption (Zinyama, 2018). Rebuilding trust is critical to Zimbabwe in improving governance, public administration, combating corruption and reconnecting the citizens with the government. There is, therefore, an urgent need to come up with viable blue-prints for good governance Zimbabwe. This paper seeks to interrogate issues of politics and ethics because they have a bearing on the realisation or lack thereof of the Zimbabwean Vision 2030. In addition, this paper comes at a time when the government has already outlined its vision which clearly stipulates that Zimbabwe must become an upper-middle-class economy by 2030. This paper, therefore, seeks to provide strategies that can be employed to realise Zimbabwean Vision 2030.

The paper has four sections. Section one presents the background and conceptual framework where ‘man in the state of nature’, corruption, politics and ethics are operationalised. Section two focuses on the Zimbabwean anti-corruption system and experiences. In addition, the anti-corruption initiatives are presented and discussed such as public finance management reforms, ease of doing business reforms, electronic-government, anti-corruption structures, biometric measures, and lifestyle audits and public declaration of assets among others. Section three discusses the solutions and recommendations that can be employed to improve the institutionalisation process for anti-corruption such as ease of doing business, corruption and public procurement, reforming public sector management and fiancé, ending impunity, citizen empowerment, whistleblower protection and creating strong institutions among others. The last section presents the conclusion which concludes that corruption is a multifaceted issue that pervades all aspects of the public and private sectors, and cannot be addressed through legal reforms and enforcements without tackling the underlying causes and incentives for corruption. Hence, the need for a holistic revision of the national integrity system.

BACKGROUND AND CONCEPTUAL FRAMEWORK

It is the argument of this paper that corruption can be effectively contained after comprehending the human behaviour. It is important to start by discussing the state of nature of human beings in order to deduce the ethical and political perspectives that form the environment upon which corruption thrives.

Man in the State of Nature

One of the questions that preoccupy the philosophers and political thinkers' minds is: what is man – angelic or devilish? Assuming that there had been not powerful institution to administer and control human beings, in such a case what would have been people's behaviour toward one another? Would they have mutually respected and observed their rights or would they, like wolves, have fallen on and torn one another apart? Any sort of answer to this question necessitates the existence of a specific political and educational system. If one says that human being is intrinsically iniquitous, in that case he/she inevitably needs to perpetually control individuals. If one declares that human beings are innately angelic, it follows that he/she removes all restrictions and limitations and set them free. In order to comprehend corruption well, there is need to interrogate the views of two great thinkers. Hence, this section engages the viewpoints of Thomas Hobbes especially his famous statement, "Man is the wolf of other man." Subsequently, the section explores the views of Jean-Jacques Rousseau regarding human nature.

Thomas Hobbes' Views

Thomas Hobbes (1588-1679) is an English political philosopher who sought to apply rational principles to the study of human nature. One of the key concepts of Hobbes is the expression, "*Man in the State of Nature*", that is a hypothetical state wherein there is no political institution and administrative organisation existing in the society, the people being left to their own business and to do whatever they like. Since the instinct of love and defense of one's self is very strong in everybody, the people would be at each other's throats and would destroy one another. In the state of nature in which everybody is his/her master, one is at odds with the others concerning the nomenclature of things, and it is these differences that give rise to disputes and conflicts. Life in such a society is very thorny, painstaking and hazardous; in the words of Hobbes it is "solitary, poor, nasty, brutish and short." According to Hobbes, in no way is this kind of living to one's benefit and advantage. Therefore, people have to choose a person and give him/her adequate and total prerogatives to preserve security and so prevent individuals from assailing one another.

Such an approach to human beings draws Hobbes toward supreme monarchical rule, completely centralised authority (Leviathan) and the creation of powerful and commanding supervisory organisations, drags him completely away from populism and makes him conclude, thus: "It is, therefore, clear that so long as there is no government over the people to compel them to obey, they will exist in a state, which they have named 'a state of war', this war pitting individual against another" (Shumin, et al. 1999:84).

Hobbes' views on ethics, human beings and politics are highly controversial. In Hobbes' view, human beings are materialistic and pessimistic and their actions are motivated solely by self-interest. This is precisely the definition of corruption by World Bank, International Transparency, International Monetary Fund, United Nations Development Programme and many countries. It gives impetus to the creation of strong institutions to regulate human behaviour for the benefit of the majority. Thus, a state's stability

is only guaranteed by a sovereign authority, to which citizens renounce their rights. Leviathan (1651), his most celebrated work, expresses these views. Hobbes argues that for the appearance of morality in society, there is need of a centralised and resolute authority that maintains and promotes public morality. In the case of Zimbabwe, the National Integrity System comprising the executive, the legislature, the judiciary, Zimbabwe Ant-Corruption Commission (ZACC), civil society organisations, the media, Auditor-General's office, National Prosecuting Authority (NPA), political parties, Zimbabwe Republic Police (ZRP), Zimbabwe Electoral Commission (ZEC), Civil Service Commission, the Zimbabwe Human Rights Commission (ZHRC); and private sector among others are created by the government to ensure that the behaviour of Zimbabweans displays the highest degree of morality and ethical virtues. From this discussion it is clear that corruption addressed through legal reforms and enforcements targeted at the underlying causes and incentives for corruption and reducing payoffs to the individuals. Good corporate governance in both the private and public sectors are crucial. Strong governance arrangements in boards and in the running of institutions ensure that there are good checks and balances on the executive to the extent that executives can be held accountable for their actions. Weak institutions for corporate governance not only result in inefficiencies, but also encourage corrupt practices that feed into the society.

In the political realm, Hobbes is anti-democracy and a partisan of absolute monarchy. This is not a good assumption in dealing with corruption because it forbids a competitive political environment. Political accountability generally refers to the degree of responsiveness of ruling parties to the citizenry on the one hand, and to the constraints placed on the actions of public and political leaders on the other. Corrupt actions of both political and public leaders cascade down to civil society and weaken anti-corruption initiatives. Political accountability is likely to improve where there is reasonable degree of free and fair competition in the political and electoral processes. In Zimbabwe, the ruling party does everything to kill off opposition parties, and extant opposition parties, are either weak or disorganised. Further, the use of state resources for political purposes should be banned and the cost of party politics lowered through providing transparent subsidies. All these efforts should be backed by strict sanctions on parties that do not comply and declarations of assets and conduct by leaders should be strictly monitored. These proposals are anathema to Hobbes's Leviathan proposal.

Jean-Jacques Rousseau's View

Jean-Jacques Rousseau (1712-78) was a Swiss-born French writer, philosopher and political theorist. Rousseau first gained fame from his essay *Discourse on the Sciences and the Arts* (1750), an attack on the arts as a source for the increased wealth of the rich and an instrument of propaganda. In his *Discourse on Inequality* (1755), Rousseau professes the equality and goodness of 'natural man' and asserts that the golden age of humanity occurred before the formation of society, which bred competition and the corrupting influences of property, commerce, science and agriculture. The *Social Contract* (1762), influential during the French Revolution, claims that when human beings formed a social contract to live in society, they delegated authority to a government. However, they retain sovereignty and the power to withdraw that authority when necessary.

Jean-Jacques Rousseau's views are diametrically in opposition to that of Hobbes. The Universal Declaration of Human Rights, on which his famous statement has left its imprint, owes him too much. In the beginning of his celebrated book, *The Social Contract*, Rousseau writes, "Man is born free; but he lives everywhere in slavery." Article 1 of the Universal Declaration of Human Rights, thus, also states, "All human beings are born free and equal in dignity and rights."

Rousseau believes that man in the state of nature is decent, well-mannered and peace-loving and that it is the society which corrupts him. In his opinion, the debasement of man commences when he enjoins the civil society and relinquishes his own freedom. A human being left to himself/herself would never resort to attacking others and waging war against anybody:

Man is by nature amiable and timid, he runs away from least danger. He acquires a pugnacious temperament by virtue of habits and experience. Pride, interest, prejudgments, vengeance and all yearnings that can draw man to welcome the risk of death do not exist in nature. It is only when man enters human society that the thought of assaulting others enters his mind. After becoming a citizen he changes into a soldier. Therefore, man, by nature, has no inclination to wage war against his fellow human beings.

So long as man lives in the lap of nature and is not a captive of society, he/she is in harmony and intimacy with all the constituent parts of nature. His/her needs are limited and can easily be met. Neither is there any sign of avarice and covetousness, nor envy and the killing of one another. However, it is regrettable that this state of affairs does not last long. It is not clear why man abandons this comfort and serenity, and decides to establish a human society. This act is tantamount to forfeiting one's own natural freedom and destroying one's own pure nature and natural disposition, for "it is the society that corrupts and defiles beings...the more human beings gather together, to the same extent will they be further corrupted."

The source of human wars and conflicts is the desire to own, which in turn is an offshoot of society. It is this longing for possession that drives human beings to corruption. Gleaned from Rousseau's perspectives regarding the behaviour of human beings is that it is fundamentally the environment that toxifies and moulds the behaviour. For instance, a disabling political and socio-economic environment for economic growth and equitable distribution of resources promotes corruption. Thus, the following components of the environment causes corruption: wage incentives (inadequate pay, fringe benefits); sociological (custom, family pressures on government officials and ethnicity) cultural factors; leadership. It is also noteworthy that poverty, political instability and other societal forces increase public servant's susceptibility to corruption, especially when public officials feel that opportunities may vanish following defeat at polls or when their kinsmen place large demands on them. The trillion advice from Rousseau is that the environment promotes corruption, thus, any effective comprehensive anti-corruption strategy should pay attention to the human agency.

Corruption

This paper does not seek to define corruption. However, corruption is one of the most chronic macro-economic problems confronting most African countries in the twenty-first century (Zinyama, 2018). It has been considered the root cause of the various economic and political crises that have plagued the African region, and continues to aggravate not only the problem of underdevelopment of each country, but also that of abject poverty of the citizenry. Corruption takes place in the form of bribery, kickbacks, commissions, or other benefits.

Thus, corruption is in its simplest terms, the abuse of power for personal gain. Nyantakyi (2002:10) defines corruption as:

an act or a conduct of dishonesty committed by way of omission or commission which is intended to implicitly or explicitly influence, deviate from and alter the just behaviour and accepted societal propriety in order to satisfy one's selfish and/or parochial interests.

Corruption is a highly complex phenomenon. The parties involved leave very little telltale in the form of irrefutable hard evidence. Most of the corruption takes place in an informal manner and under the dark cover of isolated contacts. At times it does not even require a spoken word. Mere eye contact can establish a relationship of corruption. Societal moral virtues and ethics are eroded consequently.

Politics

Politics is the struggle for the acquisition of power. The nature of power corrupts people. A person who feels powerful rebels and forgets himself/herself, as well as his/her being a servant of God and nourishes an illusion of godhood. Power is prone to excesses and does not stop at a certain point. For instance, Hitler was ready to annihilate the whole of humanity and make himself stay in power in Germany. Thus, the evils of power should be identified and their emergence prevented. Francois Mitterrand, the former president of France who spoke about the danger of power, regards the solution to lie in controlling and guiding it, and in his words, in acquiring anti-power: “if the one who is in power...is of sound reasoning, he will search for anti-power” (Limba, 2008; Black, 2016). This anti-power prevents the excesses of power as well as the corruption of the power-holder. Anti-power within a person is called conscience. “Some of my powers only rely on my conscience” (Mitterrand cited by Limba, 2008). But one should not content himself with this anti-power; rather, “the society, likewise, ought to bring into existence anti-powers within the framework of its institutions” (Ibid.). In other words, power should be controlled through the mechanisms of distribution of power, the partnership of the people, and the supervisory bodies. Examples of anti-power mechanisms include the national integrity system as shown in Table 1 below.

Table 1. A National Integrity System

Institutional Pillar	Corresponding Core Rules/Practices
Executive	Conflict of interest rules
Legislature/Parliament	Fair elections
Public Accounts Committee of Legislature	Power to question senior officials
Auditor General	Public Reporting
Public Service	Public Service Ethics
Judiciary	Independence
Media	Access to information
Civil society	Freedom of Speech
Anti-corruption/Watchdog Agencies	Enforceable and enforced laws
Private Sector	Competition policy including procurement rules
International community	Effective Mutual legal/judicial assistance

Source: Transparency International, 2000

Nevertheless, between these two types of anti-power, it seems that the internal anti-power has primacy. Thus, the true power is the one which people exercise on their own. This is where politics and ethics are imperative. Thus, not only politics could reduce corruption, but should be, ethical. Even Machiavelli, who stresses so strongly on the independence of politics from ethics, believes that immorality is more dependent on the type of government, not on the principle of politics. Ethics and politics is the unity of the two. Lying, oppression, injustice, mischief, corruption and dishonest both in the individual and social spheres are bad.

Ethics

The pivot of ethics is truth and right while the motive of politics is interests and benefits. Ethics demands that people tell the truth even though it is against them, not to do injustice, not to take people as instruments, to be advocates of justice all the time, not to lie, to abstain from deception, not to conceal the truths. This is where politics necessitates the abandonment of some principles of ethics. Politics is nothing but arena for the expansion and preservation of power, which cannot be realised without sacrificing the principles of ethics. The problem is that if the people realise that their leaders are not behaving morally, they also wash their hands of ethics. A government which permits itself to commit injustice and deceive the people cannot expect justice and truthfulness from them. A citizen, who realises that his/her sovereign govern corrupts, fills his/her tax declaration form with lies and give wrong information. So, the government is in need of ethics. Therefore, if ethics is needed for the people, it can only be kept when the people feel that the government is also faithful and committed to the principles of ethics. Otherwise, there is no guarantee for the survival of ethics in society, and in turn, survival of the government. That is why even Machiavelli stresses on the need for the government to behave morally. Machiavelli states, “Just as good law is needed for the preservation of good morality, good morality is also necessary for the observance of law.” Thus, no government, no matter how powerful and versatile it may be, can exact obedience from the people only through police methods and by strengthening its own security system it has no alternative but to benefit from ethics and its promotion. The truth of the matter is that a moral person is the very same social and political person. Therefore, it is naïve to think that politics can be regarded as separate from ethics. In his book entitled, *Trust: the Social Virtues and the Creation of Prosperity*, the Japanese-American thinker, Francis Fukuyama (1995), points to the issue on the legitimacy crisis of the American system and regards it as caused by the negligence of the society’s leaders of the principles of ethics and their fraudulent conduct in political affairs. Deceitful conduct, moral disgrace and scandals such as Watergate provided the grounds for the mistrust of the people as regards the moral conduct of the leaders.

Synthesis of State of Nature, Politics, Ethics and Corruption

Doubtlessly, Hobbes and Rousseau’s traditions and perspectives on corruption, politics and ethics possess a part of the truth. If human beings are left to themselves and no law or moral principle controls them, certainly egoism would sway them to compete with and, finally, obliterate one another. Apparently, the cynical outlook of Hobbes is more in consonance with reality than the positive view of Rousseau. Thus, a holistic comprehensive anti-corruption strategy should not only consider laws and institutions for combating corruption but also environmental, sociological and human factors that mould and form behaviour.

THE ZIMBABWEAN SYSTEM

The Zimbabwe government has been pursuing political and economic reforms since the late 1980s in a bid to promote economic growth, reduce poverty and encourage popular participation and good governance. However, corruption assessments indicate that outcomes have been far from satisfactory (Zinyama, 2018). A 2017 study by Transparency International Zimbabwe reveals that Zimbabwe has a culture of corruption. The Global Competitiveness Report (2016-2017) identifies corruption as the third most problematic factor for doing business in Zimbabwe. The 2016 World Bank Enterprise Survey, states that 38.3% of the firms that took part in the survey identified corruption as a major constraint of doing business in Zimbabwe. A separate Global Financial Integrity report states that the country lost an estimated US\$12 billion to corruption involving smuggling and illicit financial outflows between 1980 and 2010.

The 2019 Auditor-General Report reveals that the Government of Zimbabwe overspent by more than \$2 billion without seeking parliamentary approval, while various departments and local authorities flagrantly flouted accounting procedures. According to the Zimbabwe 2018 budget, the government was supposed to spend \$4, 6 billion from the consolidated account, but ended up blowing \$7, 1 billion. At the time of writing this paper, the excess expenditure was still to be regularised in terms of Section 307 of the Constitution. The 2019 Auditor-General Report for the financial year 2018 exposed a litany of gross accounting malpractices by government departments, parastatals and local authorities that gobbled millions of dollars. The excerpts of the report are summarised in Table 1.

Table 2. 2019 Auditor-General Reports' Excerpts

The Zimbabwe Electrification Transmission and Distribution Company (ZETDC) - a unit of the Zimbabwe Electricity Supply Authority (ZESA) - has not yet taken delivery of transformers nine years after making a US\$4,9 million payment to Pito Investments. The same contractor was paid in advance an amount of US\$561 935 by the Zimbabwe Power Company (ZPC) in 2016 and has not delivered. In addition, ZPC paid R196 064 in 2016 to York International for gas that has not been received. ZETDC was previously advised by Auditor General to pursue the delivery of goods and services already paid for but did not implement the recommendations. The case mirrors the rot in most public entities as they continue haemorrhaging millions of dollars from botched deals. Grain Marketing Board (GMB) was flagged by Auditor General (AG) report for a similar deal, where an advance payment of \$1 million in 2016 for goods and services that were never delivered. Expenses worth more than \$14 million were not accounted for by Air Zimbabwe. The company failed to provide supporting documentation for operating expenses amounting to \$13 million and petty cash expenditure amounting to \$654 587. The airline had an unexplained suspense balance of \$27 million. The company has not accounted for all aircraft. The Auditor-General flagged a case involving examination management body Zimbabwe Schools Examinations Council (ZIMSEC), which continues to outsource printing services after splashing US\$3 million on its own printing press in 2016. The parastatal spent US\$2,1 million for printing the 2017 June and November examination papers. The money splurged on outsourcing the service was significantly more than the US\$1, 3 million needed to commission its own printing press. The AG report noted that the printing press was installed in 2018 after a part payment of US\$3 million, leaving a balance of US\$1,3 million. Due to the outstanding balance, the ZIMSEC could not commission the printing press. As a result, the council had to outsource the printing of Ordinary Level June and November 2017 examination papers at a cost of \$251 367 and \$1 million, respectively. National Social Security Authority (NSSA) has questionable investments made against financial advice. More than \$80 million was invested in troubled banks. NSSA invested in Treasury Bills (TBs) amounting to \$20 million with Metbank during the 2018 contrary to the Authority's risk and management recommendation. The AG report indicates malpractices at the Pig Industry Board, Civil Aviation Authority of Zimbabwe, Minerals Marketing Corporation of Zimbabwe and Zimbabwe United Passengers Company. Local authorities are similarly in the eye of a storm for lack of transparency and accountability. As at 30 May 2019, only three out of 92 local authorities had audited financial statements. For example, the Auditor-General report criticises the Harare City Council for wasting ratepayers money by hiring equipment for \$222 950 instead of repairing its own at a cost of \$159 969. The asphalt plant which mixes tar was not functional. It required \$159 969 to bring it back into service. The council in turn sought services from private companies for tar mixing and in the process incurred \$222 950. It's difficult to comprehend the economic rationale of the council's decision. Local authorities in Gweru and Mutare were rapped for the same practice. Similarly, on the welfare of citizens, the Ministry of Labour and Social Welfare did not disburse Harmonised Social Cash Transfer (HSCT) to vulnerable households to reduce household poverty, protect and enhance the livelihoods of all vulnerable children so that they refrain from risky coping strategies such as child labour and early marriages. The \$9 759 722 earmarked for HSCT programme was recalled by the Treasury on December 31, 2018 after realising that no disbursements were made, thus, depriving the beneficiaries of a better standard of living. Source: Government of Zimbabwe, (2019)

The AG report shows that there is corruption in almost all key sectors of Zimbabwe ranging from local government, public health, state entities and education among others. Governance issues, such as absence of good stewardship over public resources, have continued to dominate corruption. The systems of corruption are still largely in place. There has, however, been a variety of efforts. Among the important measures instituted to combat corruption include:

Public Finance Management Reforms

Zimbabwe has stringent statutory controls on the management and expenditure of public finances, in an effort to control public expenditure and to minimise corruption, waste and inefficiency. Such measures include independent Auditor General. The Parliament has Public Accounts Committee to exercise the oversight of executive and its budgets. The review of policies with a view to promoting transparency and accountability such as Public Entities Corporate Governance; Public Procurement and Disposal of Public Assets; Money Laundering and Proceeds of Crime; and whistleblower.

Public Finance Management (Treasury Instructions) 144 of 2019 empowers the Treasury to supervise the Consolidated Revenue Fund (CRF), charges upon CRF, appropriations from the CRF, limits of State borrowing, public debt and state guarantees, safeguards of public funds and properties, among ten (10) of the key deliverables. Furthermore, the government is no longer responsible for costs incurred when officers on official duty travel with spouses, children or dependants.

The government has enacted the Public Procurement and Disposal of Public Assets Act and came into effect on 1 January 2018. The new Act repealed the Procurement Act. The Public Procurement Act regulates the procurement cycle from procurement planning, approaches to the market, evaluation and award of tenders, contract management and disposal of assets. Specific accomplishments in modernising the public procurement system in Zimbabwe include the following:

- adoption of modern Public Procurement Regulations that define the legal, institutional, and procedural framework. The Public Procurement and Disposal of Public Assets Act came into force on 1 January 2018;
- establishment of the Procurement Regulatory Authority of Zimbabwe (PRAZ) and Board appointed in January 2018;
- development and promulgation of Procurement Regulations, Standard Bidding Documents (SBDs) and Guidelines;
- development of Training of Trainers course materials. Roll out of this training program is underway whilst a professionalization and certification module is being developed; and
- development of Electronic Government Procurement (e-GP) strategy.

Regulatory and Ease of Doing Business Reforms

The ease of doing business index ranks countries against each other based on how the regulatory environment is conducive to business operations and stronger protections of property rights. Economies with a high rank (1 to 20) have simple and friendly regulations for businesses. The 2019 Ease of Doing Business reforms are aimed at ensuring that investment promotion and economic development targets set in the Transitional Stabilisation Programme are achieved by focussing on creating a favourable climate for companies to operate in line with Vision 2030 for an upper middle-income economy. In 2015, Zimbabwe was ranked at 171 out of 189 countries and in 2016 it shifted to number 161 out of 190 countries. The

rank of Zimbabwe improved to 155 in 2018 from 159 in 2017 (Bank 2019). Currently, it takes 61 days to open a business in Zimbabwe, with all the requisite licences. The goal is 50 days.

Zimbabwe is far too bureaucratic and there are too many government agencies seeking to tax local businesses. However, Zimbabwe is commendable for setting up a fully-fledged Zimbabwe Investment and Development Agency (ZIDA) that processes investment approvals within a day. The ZIDA Act seeks to establish One Stop Investment Services centre, obliges investors to respect the country's laws and to promote good corporate social responsibility within the spheres of their operations. The Act also seeks to harmonise all investment laws in an effort to reduce bureaucracy and cut the cost of doing business in Zimbabwe and, thus, repealing the Zimbabwe Investment Authority Act, the Special Economic Zones Act and the Joint Ventures Act. The ZIDA Act also seeks to provide assurance to investors about the country's commitment to property rights, thereby, stimulating economic development in line with the goals in Vision 2030.

Anti-Corruption Structures

There was establishment of Special Anti-Corruption Unit (SACU) in 2018 housed in the Office of the President and Cabinet to improve efficiency in the fight against all forms of graft and to strengthen the effectiveness of national mechanisms for the prevention of corruption. The SACU collaborates with the Zimbabwe Anti-Corruption Commission (ZACC) and the Zimbabwe Republic Police (ZRP) and other such institutions in the fight against corruption, assists ZACC and other investigative agencies of the State in the perusal and consideration of corruption dockets, subject to the issuance of authority to Prosecute by the Prosecutor General, to prosecute corruption cases referred to the National Prosecuting Authority by investigative agencies (Maodza (2019).

In a bid to expeditiously deal with graft, the Judicial Service Commission is setting up anti-corruption courts in all the country's 10 provinces. Such courts have been opened in Bulawayo, Harare, Mutare and Masvingo. The setting up of the specialised courts is meant to complement government efforts to eliminate corruption by expeditiously dealing with corruption cases brought before the courts.

Zimbabwe Anti-Corruption Commission (ZACC) has now arresting powers. In terms of Section 2 (Paragraph h) of the definition of 'Peace Officer' under the Criminal Procedure and Evidence Act [Chapter 9:07] ZACC officers are included. In terms of the law, a police officer refers to any worker of the state, country or municipality, a sheriff or other public law enforcement agency, whose duties include arrest, searches and seizures, execution of crime or for the enforcement of law or orders. This means that in addition to ZACC officers' special powers to fight corruption, they arrest any person on reasonable suspicion of any other crimes having been committed. Before the instrument, ZACC would simply make recommendations for the police to arrest in terms of the Constitution.

Biometric Measures

The government through the Transitional Stabilisation Programme (TSP) seeks to rationalise its wage bill by ensuring an auditable and verifiable head-count. The previous Civil Service Audits undertaken by Government in 2011 and 2015, respectively, point to possible existence of ghost workers in the service, who are contributing to the burgeoning public service wage bill which accounts for over 90% of total revenues. Clearly, this goes against the thrust of re-orienting budget expenditures towards growth enhancing and poverty-reducing developmental programmes. To weed out these ghost workers, the government planned to introduce a biometric registration of all civil servants. The registration process was

designed to be rigorous and involved to capture data on Letter of Appointment, Academic and Professional Qualifications, National Identification Documents, Employment Code Numbers, and Biometric data, for instance, one's unique physical attributes such as fingerprints, DNA, iris and retina pattern, using ICT. The above system ensures that every person being paid by government for services rendered is properly accounted for.

Life Style Audits and Public Declaration of Assets

The government has not enforced its policy requiring public officials to declare their assets and disclose interests in transactions that form part of their public mandate. The efforts to strengthen the asset declaration system should be stepped up, in particular expanding the coverage of high-level officials and improving mechanisms for verification, sanctions and public access. There is need for a communication strategy. At the time of writing this paper; no one knows how far government has gone in implementing the Mandatory Declaration of Assets of 2018 by senior public officials. No one knows how effective the anti-corruption courts have been because there is no framework to at least monitor and communicate with citizens.

Nevertheless, these initiatives have in many instances not been successful. There still remains a compromising policy gap in how the government has responded to the problem of corruption. Among the major reasons for the failure of anti-corruption programmes are: lack of political will among decision makers; limited power and resources to accomplish reforms; overly ambitious and unrealistic promises; uncoordinated reforms; reforms that rely too much on law enforcement; reform strategies that target only low level officials and not the senior levels; reform strategies that do not deliver quick wins and reforms that are not fully institutionalised. Besides, for the most part, anti-corruption campaigns are often actually politically motivated struggles for power. The politicians in power accuse their predecessors or their opponents of corruption.

SOLUTIONS AND RECOMMENDATIONS

Ease of Doing Business

Public procurement should be an enabler of operations rather than an inhibitor. Policies and processes need to reduce the effort required to carry out best practice procurement. There is need for straightforward and consistent processes to enable:

- increased compliance by staff with policies;
- improved collaboration within the public; and
- reduction in transaction costs both for public sector organisations and suppliers to public entities.

At the heart of good governance lies effective public financial management, which is linked to effective public procurement, political accountability, effective checks and balances, and decentralisation and local participation. The various Auditor General's reports and media reportage attested to this assertion. Establish and operate appropriate systems of public procurement and management of public finances

based on transparency, competition and objective criteria. Public procurement is a big area of focus in preventing and fighting corruption. The importance of an ethical civil service is illustrated by the statistical correlation between merit and integrity, as demonstrated by the World Public Sector Report (2018) on human resources management.

Corruption and Public Procurement

Corruption is destroying Zimbabwe. Curbing corruption would deliver an additional more than US\$10 billion in tax revenues annually across the country, money that could support much needed investments in health, education and infrastructure. From drinking water to hospitals and death traps on roads, corruption is more omnipresent than God. It is time Zimbabwe borrows from other parts of the world like China, Thailand, Singapore and Malaysia where corruption is treated with highest contempt and sternest sanctions, including capital punishment. It does not mean that these countries are inhabited by angels, but the fact that severe sanctions are often applied has proven to be deterrent in many ways. Not all thieves would be caught but those caught would not be allowed to go with impunity. It is the regime of impunity in Zimbabwe that has made institutionalised corruption. It is not enough to continue to blame the system without looking at human agency.

Reforming Public Sector Management and Finance

Reforming public sector management is crucial in the fight against corruption, as opportunities and incentives for corruption are very much deeply rooted in public sector management. This means changing the way government does its business, that is, the internal management of public resources and the administration should be reformed. The financial management systems should provide online information for effective planning, monitoring and decision-making on a timely basis, in a reliable fashion. Without an integrated financial management system across government, it becomes difficult for the government to manage effectively and efficiently its financial resources and report on them promptly to oversight bodies.

In reforming the civil service, there should be consideration for establishing a meritocracy civil service that is well-remunerated, enhanced transparency and accountability in financial management processes, reformed sectoral service delivery, and deepened decentralisation with greater accountability. Officials are frequently appointed based on loyalty to the regime rather than merit, and are often susceptible to influence from the executive.

Ending Impunity

Effective law enforcement is essential to ensure that the corrupt are punished and break the cycle of impunity, or freedom from punishment or loss. Successful enforcement approaches need a strong legal framework; law enforcement branches; and an independent and effective judicial system. If laws were implemented many leaders, politicians, their families and cronies, business people, members of other professions, legislators, councillors, civil society leaders, police bosses, security officers, nurses, judges, priests, bishops, student leaders, college principals, civil servants, journalists and many more would be in jail. Law enforcement authorities still often hesitate to investigate the powerful. It is the reason good civil servants fear for their jobs and livelihoods when they speak out.

Citizen Empowerment

Strengthening citizens' demand for anti-corruption and empowering them to hold government accountable is a sustainable approach that helps to build mutual trust between citizens and government. For example, community monitoring initiatives have in some cases contributed to the detection of corruption, reduced leakages of funds, and improved the quantity and quality of public services. All citizens in their official dealings with public bodies, civil servants, and public officials, should be required to observe a number of responsibilities of good citizenship, in relation to, honesty, lawfulness, and the prevention of corruption. Citizens should also refrain from deceptive, dishonest or fraudulent conduct, and report any actual or suspected corruption, or misconduct to a proper authority, and refrain from making frivolous or vexatious complaints or demands.

Whistleblower Protection

The individuals who report wrongdoing at work are vital for exposing corrupt schemes and actions. Too often, however, whistleblowers face retaliation after bringing corruption, fraud and financial malpractice to light. These attacks are sometimes professional, sometimes legal, and can even be directed against the whistleblower's family, property or physical and mental well-being.

Strong Institutions: Strengthening Legislation and Rule of Law

The conventional wisdom says prevention is better than cure. Managing and being reactive to corruption through arrest is costly. It instils no public confidence in the state's ability to effectively curb corruption, especially, in a context where there is high level of impunity and seemingly selective application of the law regarding who is arrested on allegations of corruption and who is not. Corruption in the top echelons of the Zimbabwe government has weakened the state institutions that are meant to protect citizens. This becomes especially clear when one looks at the way organised crime networks can act with impunity all over the country. State institutions are failing to reign in the activities of such networks. Reporting or prosecuting these criminal activities is near impossible for officials of the judicial system. Citizens' faith in democratic institutions can only be restored by taking strong anti-corruption measures and following through on promises. Political leaders need to commit to concrete steps against corruption.

Strengthening Auditor-General (AG)

Fighting corruption requires a multi-pronged attack. It requires reforms, inter alia, in public administration, judiciary, public information, and cultural values of the society. Although there is a general agreement that the AG is effective in promoting transparency, governance and accountability, the AG still lacks adequate independence and resources to play this role.

The AG should try to create, with the help of government, an institutional mechanism to feed the results of audit in future economic planning. It means, before a development project or programme is finally approved, there should be a system of clearance from an independent body that the current project or programme plan does not commit the same mistakes which the past audit reports had pointed out. Unfortunately, this important step in the cycle of good governance is badly missing. The AG keeps on reporting same or similar irregularities or examples of mismanagement and embezzlement year after

year and the executive hardly takes note of that. Future projects and programmes commit the same mistakes. It is far more effective and economical to prevent the mistakes (intentional or unintentional) than to detect them and take corrective action.

Zimbabwe Anti-Corruption Commission (ZACC)

When the institution is established, and its powers are regulated, the most difficult task starts. The ZACC must be given sufficient resources to hire and to educate its employees, to purchase the necessary premises and technical equipment, and to pay an appropriate salary to its employees. Independence in the drafting and expenditure of its budget is also a basic pre-condition for its effective operation, as well as being a clear signal of the real intentions of the country establishing such an institution. The success of ZACC depends upon a seemingly trivial matter such as money, but money is proof of genuine political will. Even the best legal arrangements for the establishment of an ideally positioned anti-corruption institution undoubtedly fail without an appropriate budget allocation.

When an institution begins to operate it must strictly follow principles that are unconditionally linked to its work: objectivity, professionalism, impartiality, integrity, honesty, effectiveness and efficiency. If these principles are not followed, those who oppose the institution easily discredit its efforts and demand its restructuring or even its abolishment.

CONCLUSION

The study of corruption has provided evidence that despite several laws and institutions for combating corruption, there has been very little long-term success. Corruption still remains high in the Zimbabwean political agenda (Policy of Zero Tolerance for Corruption) but there is no clear strategy for combating it. Many anti-corruption institutions and integrity systems were established without much success. Despite this track record of abysmal performance, Zimbabwe is still willing to continue creating anti-corruption institutions, for several reasons: political considerations, conditionalities for donor assistance or because it is the most plausible alternative to espouse.

Until the implementation and enforcement of laws is improved, systemic corruption will remain a problem. However, corruption cannot be addressed through legal reforms and enforcements without tackling the underlying political and ethical fundamentals of the people. This, therefore, means that there needs to be a comprehensive corruption strategy based on human agency. Inevitably, the strategy requires reforms spanning from empowering civil society, reforming public sector management and finance, reforming the private sector to become more competitive, reforming the political environment to become more competitive and more transparent, controlling the powers of the executive and making it more transparent and accountable.

Although strong enforcement and monitoring are needed, they will have little long-term effect if the basic conditions that encourage corruption are not reduced. Consequently, it would appear that the objective of anti-corruption should be to undertake substantial improvement in the overall efficiency, fairness and legitimacy of the state, by addressing those underlying incentives to receive or pay bribes.

FUTURE RESEARCH DIRECTIONS

Insights and directions from this study recommend that a credible anti-corruption framework should incorporate the following:

- Comprehensive anti-corruption strategy (coordination among several watchdogs agencies, effective legal framework and law enforcement system, strong judicial system, high civil salaries; integrity, transparency and accountability in government operations, tolerant public attitude, political and administrative will; strategic plan).
- Corruption must be treated as a multisectoral issue not as a legal, police, political or economic issue, but a pervasive issue that cuts across society.
- Competitive political environment;
- Reforming and strengthening checks and balances on the executive;
- Corporate governance;
- Value orientation;
- Empowering the private sector and civil society;
- Effectively functioning regulatory framework; and
- Meritocratic systems backed with pay reforms incentives.

REFERENCES

- Armstrong, E. (2005). *Integrity, Transparency and Accountability in Public Administration: Recent Trends, Regional and International Developments and Emerging Issues*. UNDESA.
- City Press. (2019). *African oligarchs and their global networks of corruption*. African Investigative Publishing Collective and ZAM. <https://city-press.news24.com/News/african-oligarchs-and-their-global-networks-of-corruption>
- Coady, C. A. J. (1991). *Politics and the Problem of Dirty Hands*. In *A Companion to Ethics*. Basil: Blackwell.
- Dininio, P., & Kpundeh, S. R. A. (1999). *Handbook on Fighting Corruption*. USAID.
- Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. New York: The Free Press.
- Government of Zimbabwe. (2017). *Public Procurement and Disposal of Public Assets Act* [Chapter 22:23]. Harare: Printflow (Pvt) Ltd.
- Government of Zimbabwe. (2019). *Report of the Auditor-General for the Financial Year ended December 31, 2018 on Appropriation Accounts, Finance and Revenue Statements and fund Accounts*. Harare: Parliament of Zimbabwe.
- Government of Zimbabwe. (2019). *Statutory Instrument 144 of 2019: Public Finance Management (Treasury Instructions)*. Government Gazette.

Government of Zimbabwe. (2019). *Criminal Procedure and Evidence (Designation of Peace Officers) (Amendment) Notice, 2019 (No.3) modified Statutory Instrument 227 promulgated 1997*. Government Gazette.

Hobbes, T. (1994). *Leviathan* (E. Curley, Trans.). Hackett. [1651/1668]

Kaufmann, D. (2005). Back to Basics -10 Myths about Governance and Corruption. *Finance & Development*, 1, 66–79.

Machiavelli, N. (1988). *Prince* (Q. Skinner, Trans.). Cambridge University Press.

Maodza, T. (2018). *ED sets up special anti-corruption unit*. <https://www.herald.co.zw/ed-sets-up-special-anti-corruption-unit/>

Murwira, Z. (2019). President's advisors table anti-corruption measures. *The Chronicle*. <https://www.chronicle.co.zw/presidents-advisors-table-anti-corruption-measures/>

Nyantakyi, K. (2002). Corruption: What is it? In *Corruption and Development in Africa*. Ghana Academy of Arts and Sciences.

Shumin, G. (1999). *Rousseau Jean-Jacques: The Social Contract: Text and Context*. Academic Press.

Sithole-Matarise, E. (2017). Zimbabwe's Mnangagwa Promises Zero Tolerance in Corruption Fight. *Reuters*. <https://www.reuters.com/article/us-zimbabwe-politics/zimbabwe-mnangagwa-promises-zero-tolerance-in-corruption-fight-idUSKBNIEE25P>. Accessed 28 June 2019.

Whitton, H. (2001). *Implementing Effective Ethics Standards in Government and the Civil Service*. *Transparency International*. <https://www.oecd.org/mena/governance/35521740.pdf>

Zinyama, T. (2019). Corruption and Good Governance in Zimbabwe: The Need for a Comprehensive Public Accountability System. *Journal of Public Policy in Africa*, 5(2), 110–131.

KEY TERMS AND DEFINITIONS

Corruption: Corruption is an act of dishonesty committed by way of omission or commission which is intended to implicitly or explicitly vary the actions and accepted societal propriety in order to gratify one's egotistical interests.

Ethics: Ethics refers to what is morally right or wrong. That is socially acceptable behaviour.

National Framework: National framework refers to the constitutional, legislative, regulatory, institutional arrangements, processes and systems that underpin the governance of an organisation.

Politics: Politics is the struggle for the acquisition of power in order to decide who gets what, how, and when.

State of Nature: 'State of Nature' is an imaginary state wherein there is no political institution and administrative organisation existing in the society and the people do their own business. Since the instinct of love and defense of one's self is very strong in everybody, the people would be at each other's throats and would destroy one another.

Section 12

Government, Policy, and Crisis Management

Good Practices in E–Government Accessibility: Lessons From the European Union

12

Fernando Almeida <https://orcid.org/0000-0002-6758-4843>*Polytechnic Institute of Gaya, Portugal***José Augusto Monteiro** <https://orcid.org/0000-0001-8743-3826>*Polytechnic Institute of Gaya, Portugal*

INTRODUCTION

e-Government has assumed a growing importance in society, enabling the modernization of public administration through the use of information and communication technologies. Several studies worldwide, particularly in United States, Europe and Gulf countries, have reported the evolution of this paradigm and proposed several frameworks for its adoption (Baker, 2009; Alhomod & Shafi, 2012; Mansour, 2012). However, technology is not the only pillar that contributes to the success of an e-Government system, and the involvement of public institutions, their employees and citizens is also crucial. The aim of e-Government is to bring public administration and public services closer to citizens in order to maximize the efficiency of public services and, therefore, contribute to a more democratic and fair society.

At European level, e-Government is seen as one of the pillars in the construction of a Europe that promotes the integration and inclusion of citizens and contributes to the creation of European citizenship. The e-Government Action Plan 2016-2020 defines three main axes in its mission (EU, 2017a): (i) to modernize public administration; (ii) to achieve the digital internet market; and (iii) to engage more with citizens and businesses to deliver high quality services. It is conceivable to conclude that e-Government is seen by European institutions as an important policy to promote citizenship, to drive change in public services, to disseminate technology, to foster the integration of systems and processes, and to promote digital inclusion of citizens.

e-Government platform should be inclusive and provide the same access opportunities to all citizens. In this sense, e-Government should be accessible. Thus, the concept of accessibility in the context of Web applications is associated with the effective availability of information to all users, regardless of the technology and platforms used and the sensory and functional capacities of the user. e-Government services should contribute to the inclusion of people with special needs and provide mechanisms to help them overcome traditional physical barriers (Leist & Smith, 2014). In fact, the European Commission (EC) argues in its program of employment, social affairs & inclusion that it is fundamental that the European Union (EU) promotes the full participation of disabled people in society (EU, 2017b). EC clearly states that “disability is a right issue and not a matter of discretion” (EU, 2017b). Additionally, it is important to consider that the European population is aging, which increases the relevance of accessibility, because it is expected that older people will face greater difficulties in using technology.

DOI: 10.4018/978-1-7998-3479-3.ch104

In this sense, we intend to identify the accessibility standards that must be adopted in the development of European and National e-Government services. We perform a comparative analysis of each of them, emphasizing their benefits for the inclusion of citizens, but also presenting their current limitations. Likewise, we intend to distinguish between the existing accessibility standards that are mandatory and the good practices that are suggested to be relevant to the inclusion of people and to promote a growing adoption of online public services by all citizens. Finally, we identify and describe how the best practices are implemented and achieved in e-Government solutions, giving some examples in Europe and in different EU countries.

The manuscript is organized as follows: Initially a literature review is performed to identify the main relevant accessibility standards adopted in European and National e-Government services. Also in this section, this study discusses the main good practices in accessibility that can be found in e-Government platforms. After that, the methodology is presented, which includes an analysis of the top-4 countries in the e-Government Benchmark 2017. Four case studies are considered: (i) Malta; (ii) Denmark; (iii) Sweden; and (iv) Estonia. Subsequently, the manuscript presents some solutions and recommendations and provides insights on future research directions. Finally, the conclusions of this study are drawn.

BACKGROUND

Web accessibility starts from the assumption that the content provided to the users should be delivered and perceptible in, at least, one of three human senses: vision, hearing or touch (Anderson, Bohman, Burmeister, & Sampson-Wild, 2004). The available means to users' interaction with Web must provide an easy and quick way to access all content available, through one or more senses, founded in four principles (Anderson et al., 2004): (i) perception; (ii) operation; (iii) understanding; (iv) robustness. Web technologies had the virtue of making accessible the information, participation and even entertainment, for people who suffer some type/degree of disability. Nevertheless, the inability to operate new ubiquitous devices, due to some type/degree of disability, still exclude people from opportunities in areas such as citizenship, work or leisure (Baptista, Martins, Gonçalves, Branco, & Rocha, 2016).

A report of World Bank (2011) estimates a thousand million people in the World suffering some type/degree of disability. More recently, Baptista et al. (2016) project towards approximately 15% of the population around the world is disabled. The perception of facts, such as: (i) the rise of the global population; (ii) the rise of disabling diseases derived from old age and the rise of live expectancy; (iii) the rise of new disabling pathological phenomena, derived from modern societies; and (iv) the recognition of new forms of disabilities reveals the impact of this population in the social, economic and political plans (Priestley et al., 2016). Socially, the awareness emerges that no one can be considered immune, to someday, suffering from an incapacitating disease or a decrease in mobility, dexterity, and cognitive ability (Money, Lines, Fernando, & Elliman, 2011). In the economic plan, the need emerges to provide goods and services to improve the life quality of the people who suffer some type/degree of disability or whose physical conditions, restricts the access to services and the ability to carry out independent living tasks, such as grocery shopping, bill payment, and accessing social services. In the political plan, the consciousness of individuals who live with people with some type/degree of disability and the disabled people themselves, put pressure on governments to adopt policies to improve his own life quality (Priestley et al., 2016). Some of the barriers of Web accessibility that justifies differences among countries meet explanation in the socio-economical differences, more particularly on aspects like, connectivity and the cost to get access to the Internet, policies to encourage the adoption of the digital or lack of inclusive Websites (Andreasson & Snarud, 2014; Kubitschke, Cullen, Dolphin, Laurin, & Cederbom, 2013).

In broad sense, Europe is progressively getting conciseness of the asymmetries, the low percentage of the countries that have government websites fully accessible and efforts are made to the approximation of the laws, regulations and administrative provisions of its member states on the accessibility of websites of public sector bodies (EU, 2012).

Accessibility Standards in European and National e-Government Services

Web accessibility can be so important as the accessibility in the urban environment, buildings or people houses. Taking into account the growth of the newest accessible means in the last years of these environments, promoted by international, national and local legislation and initiatives, a parallelism with the accessibility in the Web can be made. Governments around the world create legislation to ensure equity in the access to the Web, in order to promote inclusive participation in the so-called information society. However, the legislation is not uniform from around the World, which causes different conformity requisites from country to country and from region to region, as for example, European Internet Inclusion Initiative (EIII), in Europe, and Section 508 in the United States (Money et al., 2011; Cumbie & Kar, 2016). In the case of European Union, its member states implement Web accessibility in different forms (Andreasson & Snarud, 2014; W3C, 2017). Due to distinct conceptions and implementation stages in this area, third-parties as World Wide Web Consortium (W3C) try to fill this empty space by setting standards and promoting digital inclusion on websites (W3C, 2017). Currently, W3C lists some for the governmental policies related to web accessibility. Despite the importance of accessing this information, it does not replace to check the country legislation available for each country.

The Web Accessibility Initiative (WAI) conceptualizes three pillars which contain the guidelines, specifications and success criteria to verify the conformity of the principles of accessibility (Cooper, 2016): (i) **content** - Web Content Accessibility Guidelines (WCAG); (ii) **authoring** - Authoring Tool Accessibility Guidelines (ATAG); (iii) and **users** - User Agent Accessibility Guidelines (UAAG). In the domain of mobile technologies, the Mobile Web Initiative (MWI), has created a concept to Web development for mobile technologies, named Mobile Web Best Practices (MWBP), which main purpose is to evaluate the capability of Websites accessed from mobile devices using its own browser (Baptista et al., 2016).

These standards are in the 2.0 version and were expected new developments in order to meet (Cooper, 2016): (i) new issues on accessibility; (ii) enforce articulation between standards; (iii) put more focus on users' needs. An important contribution of the standard WCAG 2.0 was the introduction of conformity levels to reflect the accomplishment of the criteria of accessibility: (i) websites on level A, means that level A criteria are in conformity; (ii) websites on level AA, means all criteria of level A and level AA are in conformity; (iii) and websites on level AAA, means that all criteria of level A, level AA e level AAA are in conformity. This novelty facilitates the process of studying documentation for the development teams in order to create websites more accessible even if they only accomplish the first or the second levels. Nevertheless, the growth of the use of mobile and ubiquitous devices, opens challenges for new simplifications. As referred by McCathieNeville (2009), differences among WCAG and MWBP and the inherent difficulty in the establishment of a mapping between them, imply that developers need to necessarily know specifications of both.

In the future, the trend points to the accomplishment of the commitments the citizens have with the state and forces the use of the electronic government instead presidential way, bringing issues about how the public administration websites are prepared to meet citizens/users needs. Thus, it is paramount that the public administration portals can respond to the needs of the users on order so they can honor their

commitments with the government. The adoption of accessibility standards by government and public administration portals becomes a critical factor of success to the satisfaction of users' needs to ensure their obligations with the state (Velleman, Nahuis, & Van der Geest, 2017).

Good Practices in Accessibility

Our understanding of the practices in the promotion of accessibility conditions to an inclusive access to the e-Government may focus on three dimensions: (i) a deep understanding of the specific needs of the disabled people; (ii) actions of the governments and authorities focused on creating conditions to a more inclusive digital society; (iii) and the update of the standards and the monitoring and certification tools.

The way that Websites deal with the visitors does not allow a differentiation of the profile of the visitors in order to perceive their needs, limitations or capabilities, namely, in terms of accessibility. The lack of information about the behavior of the different users who suffer some type/degree of disability, does not allow a deep understanding of their specific needs to reduce barriers on the access to the Web (Baptista et al., 2016). Laboratory tests only provide closed views of a small and "controlled" population. An evaluation on how disabled users interact with government Websites in real context may lead to a more accurate perception of their difficulties, needs and expectations.

In parallel, another approach to minimize the difficulties of people with disabilities has been provided through directives from a supra state level (ONU, USA, EU) that lead member states to adopt common procedures to inspire the industry to create more accessible products. These documents do not limit creativity of industry in the field of accessibility, instead, they define the orientations that should be followed to produce better and more accessible products. An example is the European Union, which directive (EU) 2016/2102 main purpose is to promote accessibility of the websites and mobile applications of public sector bodies, in all member states (Parliament & Council, 2016). The directive was proposed by the Commission in 2012 and, after a period of negotiations and updates, was approved by the European Parliament in 2016. Member states have to transpose the most relevant provisions of the directive to their national legislation until September of 2018. The directive added on the European Accessibility Act, which aims to improve the functioning of the internal European market for accessible products and services by removing barriers created by divergent legislation.

Standards and mechanisms of evaluation of the conformity play a very important role to validate and enforce, in a technical way, legislation purposes. Despite the simplification in the WCAG standards with 2.0 version, some criticism mentions difficulties in validation of the directives. The practices carry on to evaluate accessibility of the Websites is the adoption of multiple techniques, which vary from heuristics (Hermawati & Lawson, 2016) to more complex models, in some cases, called quality models (Afonso, Angélico, Lima, & Cota, 2014). European Internet Inclusion Initiative (EIII) inspired on the directive to the accessibility of public sector bodies' websites combines existing web accessibility evaluation tools and supports user testing through crowd-sourcing (Andreasson & Snarud, 2014; Parliament & Council, 2016). According to Andreasson & Snarud (2014), it combines the benefits of automated evaluation tests with manual input and aims to be the first open source WCAG 2.0 compliant tool to provide a "mixed" test approach.

Despite the importance given to the application of good practices in the design of e-government solutions, there is no specific world ranking in this field. The current literature focuses on comparing the penetration rate of e-government services in several countries around the world on multiple perspectives such as e-participation, adherence to open data formats and cyber security policies (Obi & Iwasaki, 2015). Finally, the study conducted by Meskeel (2014), which presents a list of best e-government websites in

the world, deserves special mention. This ranking is established taking into consideration the design of the websites, highlighting as important contributions the use of responsive design, the simplicity of navigation, prominent search bars, grouping content by themes, use of icons, and the availability of multiple languages.

APPROACH AND DISCUSSION

e-Government Benchmark 2017 (EU, 2017c) realizes an assessment of e-Government services in 34 European countries, which include EU28+, as well as Iceland, Norway, Montenegro, Republic of Serbia, Switzerland, and Turkey. Four top level benchmarks indicators are established according to the e-Government Action Plan: (i) user-centric government; (ii) transparent government; (iii) cross-border mobility; and (iv) key enablers. Top-4 countries in the overall e-Government Benchmark 2017 are: (i) Malta; (ii) Denmark; (iii) Sweden; and (iv) Estonia. The analysis focused precisely on these four countries that present an overall performance in the e-Government since they appear to have potential for the identification of good practices in e-Government accessibility.

The accessibility analysis carried out followed a two phase approach. The first step consisted in the application of an automatic usability analysis tool. Several tools were considered from the Web Accessibility Evaluation Tools List¹ provided by W3C. The decision was to use the WAVE² tool, which facilitates web accessibility evaluation by providing a visual representation of accessibility issues within the page and it offers a Firefox and Chrome extension. WAVE organizes the accessibility issues in terms of errors, alerts, features, structural elements, HTML5 and Accessible Rich Internet Applications (ARIA), and contrast errors. This approach allows us to find the main accessibility issues. In a second phase perform an exploratory study through a manual inspection considering the government portal of each country. Typically, the institutional government portal is the place where citizens can obtain safely and efficiently the public services they need. Adopting this approach allow us to find good accessibility practices according to the specificities of each e-Government portal.

Malta

The first case study addresses Malta's e-Government portal³. Table 1 presents an overview about the main identified accessibility issues. Although Malta ranks 1st in the overall e-Government Benchmark 2017, we found several accessibility issues, particularly structural elements. At this level, we found is-

Table 1. Accessibility issues in Malta's e-Government portal

Category	No. of issues	Description
Errors	12	Linked image missing alternative text, missing form label, empty button and empty link.
Alerts	22	Suspicious alternative text, broken same-page link, suspicious link text, redundant title text, among others.
Features	14	Alternative text and linked image with alternative text.
Structural Elements	85	Layout table and unordered list.
HTML5 and ARIA	35	Navigation issues and ARIA tabindex.
Contrast Errors	12	Very low contrast.

sues in the construction of the tables and in the creation of lists. While tables are primarily intended for the presentation of tabular information or data, they are used in this website to control page layout and formatting, which can introduce reading and navigation order issues. Additionally, the website frequently presents an unordered bulleted list. Instead, it is recommended to use ordered lists that present a group of related parallel items.

Despite the accessibility issues, numerous important good practices are also presented. In Figure 1, we can easily identify the 1st and 2nd good practice described below at the top of the page. We highlight three good practices in this portal:

- The font size of the site can be customized by the user. Three letter sizes are set for this purpose;
- The search field allows users to search for content within the site and the domain. This is very important as often the e-Government service that users are looking for is not available on the main site;
- The form element has an action attribute set to a text value which references a server-side function.

Figure 1. Malta's e-Government portal



Denmark

The second case study considers the Danish Parliament⁴. Table 2 presents an overview of the main identified accessibility issues. Like in the previous example, most of the issues found belong to the category of structural elements. However, in this portal, besides the use of unordered lists it is also pointed out an incorrect use of level headings. Headings facilitate page navigation for users of assistive technologies. They also provide semantic and visual meaning and structure to the document. In this sense, its

Table 2. Accessibility issues in Danish Parliament portal

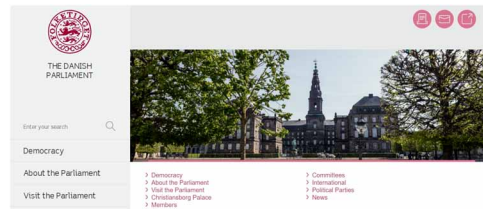
Category	No. of issues	Description
Errors	5	Empty links.
Alerts	6	Unlabeled form element with title and redundant links.
Features	6	Missing alternative text, null or empty alternative text and skip link target.
Structural Elements	25	Heading level issues and unordered list.
HTML5 and ARIA	9	Navigation issues and ARIA tabindex.
Contrast Errors	3	Very low contrast.

non-adoption in terms of `<h1>` and `<h2>` elements turns more difficult to identify the most relevant elements in the site. It is also worth mentioning the use of a symbol that identifies the main page functionalities, such as printing, e-mailing or dissemination of social networks (Figure 2). However, as these symbols are not standard, it may become difficult for users to access correctly and quickly the functions associated with these buttons.

The Danish Parliament site also presents some good accessibility practices. Among them we highlight two:

- Searched elements can be shown by relevance, date and title. The site displays 25, 50, 100, 200 entries per page (value that can be customized by the user);
- Menu is duplicated consistently. Therefore, the interaction with the user is simplified. Additionally, the color of the menu element changes when the user is on the object.

Figure 2. Danish Parliament portal



Sweden

The third case study considers the Government Offices of Sweden⁵. Table 3 presents an overview about the main identified accessibility issues. Features issues stand out most. 102 of 127 issues found are missing form labels. Consequently, it is recommended to present a properly associated form label to a screen reader user when the form control is accessed. Moreover, this approach allows the label to be clicked with the mouse to set focus on the form control. Additionally, we emphasize another accessibility issue on the main site. A Flash video presents some important events in which the Swedish government participated, but the video visualization presents two problems: (i) requires update of the Adobe Flash

Table 3. Accessibility issues in Government Offices of Sweden

Category	No. of issues	Description
Errors	0	
Alerts	15	Long alternative text, no-script element, missing accesskey and redundant title text.
Features	127	Missing alternative text, missing linked image with alternative text, missing form label, missing fieldset and skip link target.
Structural Elements	41	Heading levels issues, bad construction of ordered list and unordered list issues.
HTML5 and ARIA	66	Issues in header, navigation, main content and footer. Additionally, missing ARIA labels or descriptions.
Contrast Errors	6	Very low contrast.

Player; and (ii) the commands of the videos are in Swedish, although English has been defined as the language of interaction with the website.

Some good practices were also identified on the Government Offices of Sweden portal:

- The website is responsive and it can be properly navigated in desktop, cell phones and tablets devices. This situation is illustrated in Figure 3;
- In spite of the excess of information that is presented to the citizen on the main page, we register positively the non-use of pop-ups.

Figure 3. Government Offices of Sweden portal in a cell phone device



Estonia

The Estonia e-Services portal⁶ is our fourth case study. This is the aggregating portal of various e-Government services available in Estonia. Rather than having to hunt for a particular service on the Internet, Estonian citizens can simply visit this gateway site. The main found accessibility issues are presented in Table 4. The focal problems are in the use of colors that have very low contrast. This situation hinders navigation on the website. Additionally, there is an Estonian promotional video that is always loaded in the background. In addition to the struggle to visualize content, it also makes website loading a lot heavier, especially on devices with lower memory capacities. The organization of e-Government services is also another element that presents some vulnerability, since they are presented by their order of availability to the public, and it is not possible to organize them in any other way. Finally, the descriptive text associated with the features implemented by Estonia government is only discovered after the user has selected each of the topics.

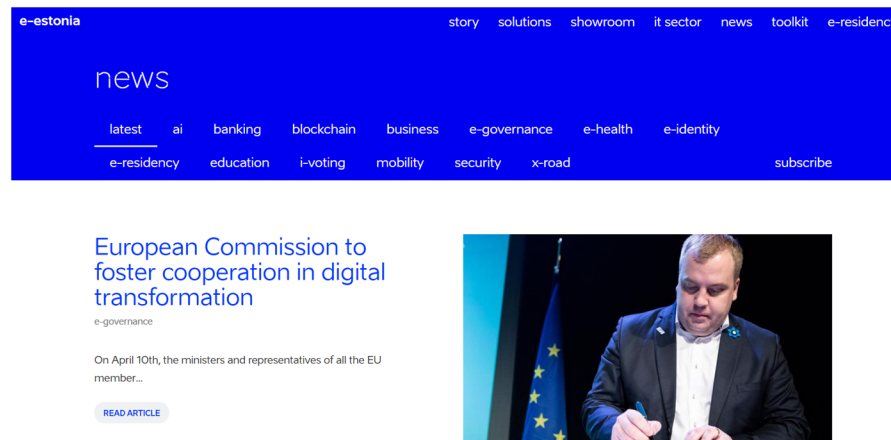
Some good practices were also identified on the Estonia e-services portal:

- There is a news subscription process that is organized by various topics according to the interests of the citizens (Figure 4);
- Elements are arranged according to the intended position of their content in the reading order of the displayed content;
- The use of submenus within menus is avoided. The only exception is the “news” menu, which access is also made possible by using the keyboard.

Table 4. Accessibility issues in Estonia e-Services portal

Category	No. of issues	Description
Errors	22	Missing form label and empty link.
Alerts	12	Redundant alternative text, skipped heading level, suspicious link text, redundant link and link to PDF document.
Features	25	Missing alternative text and null or empty alternative text.
Structural Elements	40	Missing head levels, unordered list and inline frame.
HTML5 and ARIA	9	Issues in header, navigation, search and footer. Present also ARIA tabindex issues.
Contrast Errors	134	Very low contrast.

Figure 4. Estonia e-Services portal



SOLUTIONS AND RECOMMENDATIONS

Web accessibility is a relevant social, ethical and economic issue. Increasingly there is a growing need for access to information presented on e-Government platforms. It is vital that this information is accessible to all citizens, regardless of the type of disability they have. Simultaneously, it is also important to guarantee access to e-government solutions by the older population, which is increasingly necessary with the progressive aging of the population. Therefore, it clearly states that there is a need to ensure accessibility for both people with special needs and for the elderly.

A controversial subject is the implementation of e-Government solutions that can offer an attractive design and meet accessibility standards. In this sense, the dichotomy between design and accessibility emerges. Consequently, it is fundamental to reinforce the idea that both approaches are compatible. For its realization, some recommendations are suggested, such as simplicity in structuring information, concern with the semantics of navigation that can be performed with touch screens, mouse, keyboards or other devices, concern about the readability of both text and images, and testing the e-Government solutions to measure compliance with accessibility standards before the website is available to the public.

FUTURE RESEARCH DIRECTIONS

One of the main points of future research is the development of a systematic study that performs a quantitative analysis of the accessibility levels offered by the main e-Government services in the European Union. This systematic analysis should seek to identify and perform a comparative analysis of the main accessibility vulnerabilities. It is also relevant to analyze the correlation between the e-Government performance and the degree of compliance with accessibility standards.

Another research point should focus on analyzing the impact of major accessibility vulnerabilities on the public with special needs. It is intended to explore the impact of these vulnerabilities on this public, since it is expected that not all vulnerabilities found will have the same impact and importance.

CONCLUSION

Information technologies are fundamental elements in the modernization of public administration. e-Government is a key strategic process for improving democracies and promoting digital inclusion. e-Government solutions should ensure that Web accessibility standards and good practices are implemented, so that people with special needs can access public services and can be active and inclusive agents in society. In this study, it was possible to identify a set of good accessibility practices implemented in the e-Government services of the top-5 countries in the overall e-Government Benchmark 2017. It is expected that by disseminating these good practices in these reference countries in the e-Government field its adoption and dissemination may be encouraged in other countries and e-Government services.

Despite the identified good practices, which are important elements in the integration of citizens with special needs, it was possible to conclude that guaranteeing total accessibility and meeting all the rules defined in WCAG 1.0 and WCAG 2.0 standards is still a challenge. The main difficulties found were the construction of tables to define the layout of the sites, creation of unnumbered lists, incorrect use of level heading, missing alternative text in images and levels, adoption of flash technology and use of colors that present very low contrast.

The obtained results are pertinent and have significant managerial impact for Web developers and designers of e-government platforms. The found issues serve as a reference for the implementation of e-government solutions in which the same mistakes are not made and to ensure that good accessibility practices are followed. Additionally, the good practices serve as benchmarks for their large-scale adoption in emerging e-government solutions.

REFERENCES

- Afonso, A. P., Angélico, M. J., Lima, J. R., & Cota, M. P. (2014). UsaWeb. A model for usability evaluation web interfaces. In *Proceedings of 9th Iberian Conference on the Information Systems and Technologies (CISTI)* (pp. 1–6). IEEE. 10.1109/CISTI.2014.6876952
- Alhomod, S., & Shafi, M. (2012). Best Practices in E government: A review of Some Innovative Models Proposed in Different Countries. *International Journal of Electrical & Computer Sciences*, 12(1), 1–6.

Anderson, S., Bohman, P., Burmeister, O., & Sampson-Wild, G. (2004). User needs and e-Government accessibility: The future impact of WCAG 2.0. *Lecture Notes in Computer Science*, 3196, 289–304. doi:10.1007/978-3-540-30111-0_25

Andreasson, K., & Snaprud, M. (2014). The European e-government web accessibility divide. In *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance* (pp. 391-394). ACM.

Baker, D. (2009). Advancing E-Government performance in the United States through enhanced usability benchmarks. *Government Information Quarterly*, 26(1), 82–88. doi:10.1016/j.giq.2008.01.004

Baptista, A., Martins, J., Gonçalves, R., Branco, F., & Rocha, T. (2016). Web accessibility challenges and perspectives: A systematic literature review. In *Proceedings of the 11th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. 10.1109/CISTI.2016.7521619

Cooper, M. (2016). Web accessibility guidelines for the 2020s. In *Proceedings of the 13th Web for All Conference* (p. 5). ACM.

Cumbie, B. A., & Kar, B. (2016). A study of local government website inclusiveness: The gap between e-government concept and practice. *Information Technology for Development*, 22(1), 15–35. doi:10.1080/02681102.2014.906379

EU. (2012). *Proposal for a Directive of the European Parliament and of the Council on the accessibility of public sector bodies' websites*. Retrieved 2018, April 16, from <https://ec.europa.eu/digital-single-market/en/news/proposal-directive-european-parliament-and-council-accessibility-public-sector-bodies-websites>

EU. (2017a). *European eGovernment Action Plan 2016-2020*. Retrieved 2018, April 24, from <https://ec.europa.eu/digital-single-market/en/european-egovernment-action-plan-2016-2020>

EU. (2017b). *Persons with disabilities*. Retrieved 2018, April 24, from <http://ec.europa.eu/social/main.jsp?catId=1137>

EU. (2017c). *eGovernment Benchmark 2017 Report*. Retrieved 2018, April 26, from <https://www.capgemini.com/consulting/resources/egovernment-benchmark-2017-report/>

Hermawati, S., & Lawson, G. (2016). Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied Ergonomics*, 56, 34–51. doi:10.1016/j.apergo.2015.11.016 PMID:27184309

Kubitschke, L., Cullen, K., Dolphin, C., Laurin, S., & Cederbom, A. (2013). *Study on Assessing and Promoting E-Accessibility*. Retrieved 2018, April 16, from <https://ec.europa.eu/digital-single-market/en/news/study-assessing-and-promoting-e-accessibility>

Leist, E., & Smith, D. (2014). Accessibility Issues in E-Government. In A. Kő & E. Francesconi (Eds.), *Lecture Notes in Computer Science: Vol. 8650. Electronic Government and the Information Systems Perspective*. Springer. doi:10.1007/978-3-319-10178-1_2

Mansour, A. (2012). E- Government in the Gulf Cooperation Council Countries (GCC): A Comparative Study. *Journal of the Social Sciences*, 40(1), 1–14.

McCathieNevile, C. (2009). *Relationship between Mobile Web Best Practices (MWBP) and Web Content Accessibility Guidelines (WCAG)*. Retrieved March 31, 2018 from <https://www.w3.org/TR/mwbp-wcag/>

Meskeil, D. (2014). *The Best E-gov Websites in the World*. Retrieved 2018, August 17, from <https://digital.gov/2014/12/31/the-best-e-gov-websites-in-the-world/>

Money, A., Lines, L., Fernando, S., & Elliman, A. (2011). e-Government online forms: Design guidelines for older adults in Europe. *Universal Access in the Information Society*, 10(1), 1–16. doi:10.1007/10209-010-0191-y

Obi, T., & Iwasaki, N. (2015). *A Decade of World e-Government Rankings (Global e-Governance Series)*. Ios Press.

Parliament, E., & Council, E. (2016). *Directive (EU) 2016/2102 of the European Parliament and of the Council on the accessibility of the websites and mobile applications of public sector bodies*. Retrieved April 9, 2018, from <http://data.europa.eu/eli/dir/2016/2102/oj>

Priestley, M., Stickings, M., Loja, E., Grammenos, S., Lawson, A., Waddington, L., & Fridriksdottir, B. (2016). The political participation of disabled people in Europe: Rights, accessibility and representation. *Electoral Studies*, 42, 1–9. doi:10.1016/j.electstud.2016.01.009

Velleman, E. M., Nahuis, I., & van der Geest, T. (2017). Factors explaining adoption and implementation processes for web accessibility standards within eGovernment systems and organizations. *Universal Access in the Information Society*, 16(1), 173–190. doi:10.1007/10209-015-0449-5

W3C. (2017). *Web Accessibility Laws and Policies*. Retrieved April 11, 2018, from <https://www.w3.org/WAI/Policy/W3C>

World Bank. (2011). *World report on disability: Main report*. Retrieved 2018, April 16, from <http://documents.worldbank.org/curated/en/665131468331271288/Main-report>

ADDITIONAL READING

Bélanger, F., & Carter, L. (2009). The impact of the digital divide on e-government use. *Communications of the ACM*, 52(4), 132–135. doi:10.1145/1498765.1498801

Brown, J., & Hollier, S. (2015). The challenges of Web accessibility: The technical and social aspects of a truly universal Web. *First Monday*. Retrieved 2018, August 17, from <http://firstmonday.org/ojs/index.php/fm/article/view/6165/4899>

Choudrie, J., Ghinea, G., & Songonuga, V. (2013). Silver surfers, e-government and the digital divide: An exploratory study of UK local authority websites and older citizens. *Interacting with Computers*, 25(6), 417–442. doi:10.1093/iwc/iws020

Gasova, K., & Stofkova, K. (2017). E-government as a quality improvement tool for citizens' services. *Procedia Engineering*, 192, 225–230. doi:10.1016/j.proeng.2017.06.039

Kamoun, F., & Almourad, M. (2014). Accessibility as an integral factor in e-government web site evaluation: The case of Dubai e-government. *Information Technology & People*, 27(2), 208–228. doi:10.1108/ITP-07-2013-0130

Kuzma, J. (2010). Accessibility design issues with UK e-government sites. *Government Information Quarterly*, 27(2), 141–146. doi:10.1016/j.giq.2009.10.004

Mahmood, M., Weerakkody, V., & Chen, W. (2018). The influence of transformed government on citizen trust: Insights from Bahrain. *Information Technology for Development*. Advance online publication. doi:10.1080/02681102.2018.1451980

Rubaii-Barrett, N., & Wise, L. (2008). Disability Access and E-Government. *Journal of Disability Policy Studies*, 19(1), 52–54. doi:10.1177/1044207307311533

Sarrayrith, M., & Sriram, B. (2015). Major challenges in developing a successful e-government: A review on the Sultanate of Oman. *Journal of King Saud University – Computer and Information Sciences*, 27(2), 230-235. DOI: doi:10.1016/j.jksuci.2014.04.004

Youngblood, N., & Youngblood, S. (2013). User experience and accessibility: An analysis of county web portals. *Journal of Usability Studies*, 9(1), 25–41.

KEY TERMS AND DEFINITIONS

Accessible Rich Internet Applications (ARIA): Defines a way to make web content and web applications more accessible to people with disabilities.

Flash: Popular authoring software developed by Macromedia that is used to create vector graphics-based animation programs with full-screen navigation interfaces.

HTML5: Markup language used for structuring and presenting content on the world wide web.

Pop-Ups: New window that opens in the browser when visiting a web page or accessing a specific hyperlink.

Web Content Accessibility Guidelines (WCAG): Provides a wide range of recommendations for making Web content more accessible.

ENDNOTES

¹ <https://www.w3.org/WAI/ER/tools/>

² <https://wave.webaim.org>

³ <https://www.gov.mt>

⁴ <http://www.thedanishparliament.dk>

⁵ <http://www.government.se>

⁶ <https://e-estonia.com>

National Systems of Innovation Using an Application on EU Data

George M. Korres

University of Aegean, Greece

Maria P. Michailidis

University of Nicosia, Cyprus

INTRODUCTION

Innovation is a crucial part of today's economic growth and the overall advancement of society. At the same time, this accelerator contributes to the divergence in growth between the economies that can and those that cannot afford to have innovation as a corner stone of their economy. Lack of innovation can have many sources that range from a lack of the innovative spirit of society to a lack of resources (e.g. funding, equipment, human capital). Regardless of the source, it is found that often the private sector's push is not enough to close the gap between the underdeveloped and developed economies, and it is not enough to achieve innovation-related societal goals as the private sector's orientation tends to be short term and unit-focused, while societal innovation, by design, needs to be a long-term solution for the masses. As a result, public officials design schemes that aim to both, increase involvement in innovation activities and (via the first aim) close the innovation gap.

Research and Development and technical change are directly related with industrial infrastructure, productivity effects and regional development. The term of "technological policy" indicates the national technological capabilities, and also the structure and the planning on research and development. This paper attempts to examine the role of "technological policy" and the effects on sustainable development, the implications on growth and social change, in particular.

This paper focus to examine the structure and role of technological policy and also the implications on sustainable development and social change. Particularly, this paper attempts to review the theory and the current literature of the national systems of innovation. Furthermore, it attempts to apply some statistical measurement and indices in order to estimate the effects and the implications on EU innovation systems member states. For this purpose, a presentation of data and indicators has been applied. The results indicate that there are significant differences among European countries regarding national systems of innovation and policy makers should give more emphasis on the promotion of knowledge development and diffusion, R&D activities, and R&D financing (Carvalho et al, 2015).

It is well known that the adoption and diffusion of new technologies affect the structure and the competitiveness level of the economy. The choice of technology depends upon a large number of factors: the availability of technologies, the availability of information, the availability of resources, the availability of technology itself and its capacity for successful adoption to suit the particular needs and objectives. The advanced countries in technological change have tended to put more emphasis on policies to encourage the development of research and technological activities (Freeman, 1991). Usually, technological policy should aim to create a favourable "psychological climate" for the development of research and innovations (OECD, 2010a,b).

DOI: 10.4018/978-1-7998-3479-3.ch105

Countries that innovate slower will find hard to compete in the world markets where there are many successful innovators and these countries which innovate fast may also enjoy additional gains in productivity, growth, exports, even from licensing and patent fees. The government policies in new technologies and innovations aims exactly to this point: to reinforce the technological capabilities, in order to enhance the productivity, competitiveness and economic growth of their countries. The government support is usually taken under the form of “direct” and “indirect” measures, (as for instance, different grant, loans, tax concessions, and equity capital).

FOCUS OF THE ARTICLE

Most of the scenarios on national science and technological policies have been concentrated with the supply side of the science and technological system. Therefore, the governments have to examine the benefits and the cost from technological policy and related activities.

In technological and science policy economic forecasting are required, if economic gains are to be a major component of science strategy. For instance, there is little point in developing a new technology for which there is no market because of changing economic conditions. Otherwise, there is little point in developing a technology in a country when there are good reasons to believe that another technology will be developed in another country and supersede the indigenous technology, possible even before it is developed. The governments should pay more attention in the following points: First, to deal with multiple policy objectives in the establishment of priorities including their quantification. Second, to deal with uncertainty in the ex-ante assessment of cost and benefits for the proposed government-financed programmes. Third, to compare the cost-effectiveness of government intervention with other alternatives solutions. Fourth, to identify the appropriate type of government intervention. Technological policies aim to support and to promote the new technologies through different “direct and indirect measures”. The “direct measures” usually include different subsidies, or different favourable tax treatments for research and technological activities. The “indirect measures” are carried out in the pursuit of other policy objectives (i.e. competition policy, monetary, fiscal policies etc.) and consequently affected the different research and technological activities. If there is availability of data and necessary information for research and technological activities, the safer plan is to make a separate analysis for each economic sector concerning the research and technological activities and take it into account in the perspective plans. The next step is to consider and to analyse some specific sectors that can be served as a guide for further government action. The final step is to choose the method for government action (Carvalho et al, 2015).

The experience in the most of advanced countries shows that the economic growth has been close related with that of technological growth and technological planning. The history of advanced technologically countries indicates that technology transfer has been essential contributed to industrialisation and to modernisation of the whole economy in the new industrialised countries and advanced countries. However, most of the advanced technological countries import a substantial part of the technology that they use, (as for instance happened with Japan and the other European advanced technological member states).

Policies designed to alter the rate of economic growth directly tend to focus on enhancing the technological advances and the quality of labour force. The rate of technical change is affected by research expenditures and the rate of improvement of the quality of labour force is affected by investment in human capital (such as training, and education). The investment in human capital affects positively the rate of technical change.

BACKGROUND

1. The Global Environment

The experience in the most advanced countries shows that economic growth has been closely related with technological growth and technological planning. Policy research must be undertaken to determine approaches in order to reduce costs, increase access, select more effective technologies and ensure their integration, adoption and diffusion (Acs, et al. 2017).

Within this framework, the concept of innovation policy emerged as innovation policy. Nowadays, innovation policy is regarded as a part of economic policy, and its main functions are (Marzenna 2018):

- strengthening linkages in the national innovation system;
- creating favorable conditions for the implementation of new solutions (innovation);
- fostering structural changes in the industry (e.g. changes in technology, quality improvements);
- enabling companies to reap the benefits of globalization and international cooperation.

Modern innovation policy integrates some selected elements of the policy toward science and technology as well as industrial policy in order to promote development, through an effective use of new products, services and processes by enterprises, public and private organizations as well as individual people.

Innovation policy combines many different elements related to technology, education, infrastructure, etc. Innovation policy in the EU has over 20 years of history. The evolution of the innovation concept, from the linear model, which assumes that innovation is mainly a result of research and development, to a systemic model, in which innovation is the result of interactions between people, organizations and institutions caused significant adjustments in innovation strategy and policy concepts and their implementation (Marzenna 2018).

2. A Dynamic Strategy for a Coherent National Technological Policy

The concept of national systems of innovation (NSIs) emerged in the late 1980s and started to diffuse more rapidly in the early 1990s with the seminal contributions of Lundvall (1992) and Nelson (1993). It has attracted the attention of many innovation researchers and policymakers (Amable, 2000; Edquist, 2004; Freeman, 2002; Mytelka and Smith, 2002) and has rapidly achieved broad international diffusion. However, there is therefore a need for theoretically based empirical research towards varieties of NSIs, as well as determinants of innovation processes within them (Edquist, 2004, Edquist and Hommen, 2008). The analysis of system of innovations helps us understand and explain, why the development of technology is necessary in a certain direction and at a certain rate (Freeman, 1995). The government engages itself towards innovation policy because it has been considered that innovation is a key point for the national economic growth. In order to decide how the governments should decide to promote the innovations, it is useful to know the specific context in which the national government interferes (Bergek, et al., 2008).

The first approach and definition of system of innovations is that, it is a social system that is constructed by several elements, while there is a close relationship between these elements. These elements are interacting in the production, diffusion, and economic cycles. We can define the system of innovations from a narrow view. According to the narrow definition it includes organisations and institutions that of involved in searching and exploring the new technologies (such as technological institutes, and research departments). From the other side, the broad definition follows the theoretical perspective and includes

the different parts of economic structure (such as the production system, and the marketing system). On international scene, when the large countries change the orientation of research activities, this affects the small countries. Any improvement and technological sophistication in the traditional sectors for the large countries have usually pushed the small country firms for the same sectors to follow the new technologies, (Korres, 2012, Edquist, 2004).

On the other side, the advanced technological countries are usually investing in the development of new technologies in “associated strategic industries” that aim to increase the competitiveness, the economic growth and the living standards. However, in practice different countries choose different priorities for technological and economic subjects and put emphasis, in these areas in which they believe they will have more potential in future. The advanced technological countries make their plans with more long-term scientific intensive criteria and usually the priorities that choose are based in the most expensive and high technology areas, while sometimes they carry out research in collaboration with other countries. On the other side, the small countries are forced to choose these sectors, that they easily to develop, and to compete and furthermore try to strengthen the technological basis in international competitive scene (Edquist and Hommen, 2008). The existence of specialised research and technological institutions, the amounts allocated for research and technological activities and the availability of resources in the large countries, give an additional advantage and imply that the large countries are usually leading in the trends and in the fashions for the different research and technological topics. Therefore, the small countries tend to follow the “direction” set by the leading-large countries, and sometimes they are forced to follow even if the relevant research and technological topics are not the most appropriate and necessary to cover their specific needs. It is necessary, for the small countries to pay more attention to identify the specific priorities that can exploit the advantages (such as natural and human resources, and to increase their research expenditures).

Table 1 illustrates the total R&D expenditures as a percentage of GDP for European countries for the time-period 2000-2015. This indicator provided is GERD (Gross domestic expenditure on R&D) as a percentage of GDP. “Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society and the use of this stock of knowledge to devise new applications” (OECD 1981a and OECD 1981b).

Over the last ten years, R&D intensity rose in twenty-one Member States, with the highest increases in Austria (from 2.42% in 2007 to 3.16% in 2017, or +0.74 percentage points - pp) and Belgium (from 1.84% in 2007 to 2.58% in 2017, or +0.74 pp). Conversely, R&D intensity decreased in six Member States and most strongly in Finland (-0.59 pp) and Luxembourg (-0.33 pp). In Malta, R&D intensity remained at the level of 0.55%. In 2017, the Member States of the European Union (EU) spent all together almost €320 billion on Research & Development (R&D). The R&D intensity, i.e. R&D expenditure as a percentage of GDP, stood at 2.07% in 2017, compared with 2.04% in 2016. Ten years earlier (2007), R&D intensity was 1.77% (Eurostat, 2019).

With respect to other major economies, R&D intensity in the EU was much lower than in South Korea (4.22% in 2015), Japan (3.28% in 2015) and the United States (2.76% in 2015), while it was at about the same level as in China (2.06% in 2015) and much higher than in Russia (1.1% in 2015) and Turkey (0.96%). In order to provide a stimulus to the EU’s competitiveness, an increase by 2020 of the R&D intensity to 3% in the EU is one of the five headline targets of the Europe 2020 strategy. The business enterprise sector continues to be the main sector in which R&D expenditure was spent, accounting for 66% of total R&D disbursed in 2017, followed by the higher education sector (22%), the government sector (11%) and the private non-profit sector (1%). This information on Research and Development in

Table 1. Total R&D expenditure % of GDP

	2000	2004	2005	2007	2008	2009	2010	2011	2012	2013	2014	2015
EU (28 countries)	1.77	1.75	1.74	1.77	1.84	1.93	1.93	1.97	2.01	2.03	2.04	2.03
Euro area (19 countries)	1.78	1.78	1.78	1.81	1.89	1.99	1.99	2.04	2.1	2.11	2.14	2.12
Belgium	1.92	1.81	1.78	1.84	1.92	1.99	2.05	2.16	2.36	2.44	2.46	2.45
Bulgaria	0.5	0.47	0.45	0.43	0.45	0.49	0.56	0.53	0.6	0.63	0.79	0.96
Czech Republic	1.12	1.15	1.17	1.31	1.24	1.3	1.34	1.56	1.78	1.9	1.97	1.95
Denmark	2.19	2.42	2.39	2.51	2.78	3.07	2.94	2.97	3	3.01	3.02	3.03
Germany	2.39	2.42	2.42	2.45	2.6	2.72	2.71	2.8	2.87	2.82	2.89	2.87
Estonia	0.6	0.85	0.92	1.07	1.26	1.4	1.58	2.31	2.12	1.73	1.45	1.5
Ireland	1.09	1.18	1.19	1.23	1.39	1.61	1.6	1.54	1.56	1.56	1.51	:
Greece	:	0.53	0.58	0.58	0.66	0.63	0.6	0.67	0.7	0.81	0.84	0.96
Spain	0.89	1.04	1.1	1.23	1.32	1.35	1.35	1.33	1.29	1.27	1.24	1.22
France	2.08	2.09	2.04	2.02	2.06	2.21	2.18	2.19	2.23	2.24	2.24	2.23
Croatia	:	1.03	0.86	0.79	0.88	0.84	0.74	0.75	0.75	0.82	0.79	0.85
Italy	1.01	1.05	1.05	1.13	1.16	1.22	1.22	1.21	1.27	1.31	1.38	1.33
Cyprus	0.23	0.34	0.37	0.4	0.39	0.44	0.45	0.45	0.43	0.46	0.48	0.46
Latvia	0.44	0.4	0.53	0.55	0.58	0.45	0.61	0.7	0.67	0.61	0.69	0.63
Lithuania	0.58	0.75	0.75	0.8	0.79	0.83	0.78	0.9	0.89	0.95	1.03	1.04
Luxembourg	1.57	1.62	1.59	1.61	1.64	1.71	1.51	1.47	1.28	1.31	1.28	1.31
Hungary	0.79	0.86	0.92	0.96	0.98	1.14	1.15	1.19	1.27	1.39	1.36	1.38
Malta	:	0.49	0.53	0.55	0.53	0.52	0.62	0.67	0.83	0.77	0.75	0.77
Netherlands	1.81	1.81	1.79	1.69	1.64	1.69	1.72	1.9	1.94	1.95	2	2.01
Austria	1.89	2.17	2.38	2.43	2.59	2.61	2.74	2.68	2.93	2.97	3.06	3.07
Poland	0.64	0.55	0.56	0.56	0.6	0.66	0.72	0.75	0.88	0.87	0.94	1
Portugal	0.72	0.73	0.76	1.12	1.45	1.58	1.53	1.46	1.38	1.33	1.29	1.28
Romania	0.36	0.38	0.41	0.52	0.57	0.46	0.45	0.49	0.48	0.39	0.38	0.49
Slovenia	1.36	1.37	1.41	1.42	1.63	1.82	2.06	2.42	2.58	2.6	2.38	2.21
Slovakia	0.64	0.5	0.49	0.45	0.46	0.47	0.62	0.66	0.8	0.82	0.88	1.18
Finland	3.25	3.31	3.33	3.35	3.55	3.75	3.73	3.64	3.42	3.29	3.17	2.9
Sweden	:	3.39	3.39	3.26	3.5	3.45	3.22	3.25	3.28	3.31	3.15	3.26
United Kingdom	1.64	1.55	1.57	1.63	1.64	1.7	1.68	1.68	1.61	1.66	1.68	1.7
Norway	:	1.55	1.48	1.56	1.56	1.72	1.65	1.63	1.62	1.65	1.72	1.93

Source: Eurostat Database

the EU is published by Eurostat, the statistical office of the European Union. R&D is a major driver of innovation, and R&D expenditure and intensity are two of the key indicators used to monitor resources devoted to science and technology worldwide (Eurostat, 2019).

On the European Union level, the most recent completed program was the Seventh Framework Program (which was complemented with the Competitiveness and Innovation Framework Program).

Table 2. Turnover from innovation % of total turnover

	2004	2006	2008	2010	2012
EU (28 countries)	:	:	:	:	11.9
EU (27 countries)	13.7	13.4	13.3	12.8	:
Belgium	12.9	9.3	9.5	12.4	11.2
Bulgaria	12.5	10.3	14.2	7.6	4.2
Czech Republic	15.5	14.7	18.7	15.3	13.4
Denmark	11	7.8	11.4	15	13.9
Germany	17.6	19.2	17.4	15.5	13
Estonia	11.9	13.7	10.2	12.3	7.8
Ireland	10.1	12.6	11	9.3	:
Greece	11	25.7	:	:	11.8
Spain	13.8	15.9	15.9	19	14.3
France	11.7	:	13.2	11.3	13.5
Croatia	:	13	14.4	10.5	10
Italy	11.9	9.1	11.8	14.9	11
Cyprus	5.6	12.3	16.1	14.7	11.4
Latvia	5.1	3.4	5.9	3.1	5
Lithuania	9.7	12.4	9.6	6.6	5.5
Luxembourg	15.6	12.5	8.9	8.3	7.9
Hungary	7	10.5	16.4	13.7	9.7
Malta	22.2	28.6	15.2	7.4	10.2
Netherlands	8.4	10.9	8.9	10.4	11.8
Austria	10.6	13.6	11.2	11.9	9.8
Poland	13.5	10.1	9.8	8	6.3
Portugal	10	13.3	15.6	14.4	12.4
Romania	16.6	18.5	14.9	14.3	5.4
Slovenia	14.3	13.3	16.3	10.6	10.5
Slovakia	19.2	16.7	15.8	23.3	19.6
Finland	14.9	15.7	15.6	15.3	11.1
Sweden	13.4	15	9.2	8.4	6.1
United Kingdom	13.9	8.5	7.3	5.2	14.1
Norway	7.2	4.8	4.6	6.1	5.2

Source: Eurostat Database.

Today, the Horizon 2020 is the chief innovation-stimulating program from the European Union, while the ninth framework program is being constructed (Marzenna, 2018). This picture has been reflected in the Innovation Union initiative, which was launched in 2010 as a flagship project of the Europe 2020 strategy. The Innovation Union brought a new approach to innovation policy shifting its focus from pure research to research and innovation, which address societal challenges. Innovation has been regarded as a key factor to achieve smart, sustainable and inclusive growth (European Commission, 2010).

This comparative analysis confirmed that in the period of 2010- 2017 the EU maintained its strong innovative position in the world economy and diminished the innovation gap towards the US and Japan, but was not able to reduce the distance towards South Korea. At the same time China started to catch up with innovation leaders, including the EU. Other emerging BRICS economies were not that successful in closing the innovation gap dividing them from the EU, however, some progress can be seen in the case of Russia and South Africa. These new trends need to be taken into account when setting new objectives for a future innovation strategy of the EU (Marzenna, 2018).

Table 2 illustrates the turnover from innovation as a percentage of total turnover. This indicator is defined as the ratio of turnover from products new to the enterprise and new to the market as a % of total turnover, (OECD 2010a,b). It is based on the Community innovation survey and covers at least all enterprises with 10 or more employees. An innovation is a new or significantly improved product (good or service) introduced to the market or the introduction within an enterprise of a new or significantly improved process.

This indicator is defined as the ratio of turnover from products new to the enterprise and new to the market as a % of total turnover. It is based on the Community innovation survey and covers at least all enterprises with 10 or more employees. The main conclusion that we can draw from this work and its main contribution to economic research and for companies, is that private R&D is the engine of innovation, thus, the EU should invest more in stimulating private R&D, with tax policies and other benefits, to promote investment in innovation (Carvalho et al, 2015):

- EU should continue to promote policies to support innovation in companies
- EU must innovate, finding new ways to encourage private initiative in the field of innovation seeking, thereby, to increase the number of innovative enterprises
- EU should continue and, if possible, increase public funding for innovation, as this is one of the engines of innovation.

There are a lot of proposals and strategies that can be followed and developed from the small size and weak technologically countries. Among them, an interesting view suggesting that they should develop some “new high technology areas” that may have not been developed and expertised, but the perspective potential economic benefits may be greater and they can develop a comparable advantage in the future, (Filippetti, A. & Archibugi, D., 2011). Another interesting view for the small countries suggesting the relevant strategies that are related to specialisation in some certain points for research and technological activities, in order to make an appropriate use of available resources. (Nunes et al, 2013).

Most of the scenarios on national science and technological policies have been concentrated with the supply side of the science and technological system. Therefore, the governments have to examine the benefits and the cost from technological policy and related activities. Usually, a main objective of technology policy is to increase and to enhance the use of new technologies. These technologies can be derived either from abroad or from domestic innovators and can be used domestically owned, or, in owned foreign firms. Although different countries can be choosing to develop the same kind of technologies, the policies that usually follow can differ considerably. Technological policies are based on the role of government’s intervention and are the relevant chosen priorities (such as the financial support). The divergence of national policies emergence from differences between national systems and varying views in relation to the role of the government (Edquist and Hommen, 2008). According to the different government’s policies some countries give unfair advantage to their companies in the international competition affecting the development of research activities and the new technologies. The way in

which priorities are combined and formulated in practice can vary according to the level of priority. For instance, a number of priorities that are not scientific and technological in a “strict sense” nonetheless have a considerable impact on the scale of science and technology.

SOLUTIONS AND RECOMMENDATIONS

A main objective of technology policy is to increase and to enhance the use of new technologies. These technologies can be derived either from abroad or from domestic innovators and can be used domestically owned, or, in owned foreign firms. Although different countries can be choosing to develop the same kind of technologies, the policies that usually follow can differ considerably. Technological policies are based on the role of government’s intervention and are the relevant chosen priorities (such as the financial support). The divergence of national policies emerge from differences between national systems and varying views in relation to the role of the government. According to the different government’s policies some countries give unfair advantage to their companies in the international competition affecting the development of research activities and the new technologies. The way in which priorities are combined and formulated in practice can vary according to the level of priority (Korres G. 2012, Korres G., Tsombanoglou G. 2003).

FUTURE RESEARCH DIRECTIONS

Regarding future research, it could focus on human resources, and also to examine the manner in which these resources are used. The role of universities and research institutes is an important source that can be substantial to the radical change in technological opportunities and infrastructure. The educational institutions can contribute to the introduction and diffusion of new techniques in different sectors (Korres G. 2012).

CONCLUSION

New technologies imply some direct and indirect effects or more specifically some micro effects (such as firms, and organisations) and macro effects (such as inter and intra-industrial and moreover regional effects) for the whole economy. New technologies play an important role to sectoral productivity, overall growth, employment, modernization, industrialisation, socioeconomic infrastructure and to competitiveness of a country. The principal effects for technological policy can be distinguished in demand and supply sides.

A stable macro-economic environment that encourages investment in the creation and adoption of new technologies is an important prerequisite. More important however, are the micro-economic policies that induce firms to share information, develop absorptive capacity and increase rates of adoption of new technologies, either directly (through subsidies, and financial schemes) or indirectly through alteration of the institutional and regulatory environment, (Korres G. 2012). Small countries are likely to need a more comprehensive and oriented policy of co-operative innovative effort, in order to develop their future capabilities and to make the necessary choice for technological priorities, (Korres, 2012).

The various objectives of technological policy may be subsumed under five headings:

- First, to improve the efficiency of the transfer of technology from foreign suppliers to the local users.
- Second, to increase the efficiency of operation of technology.
- Third, to strengthen the industrial base.
- Fourth, to develop the indigenous technological capability.
- Fifth, to smooth adjustment forced by new technologies.

In addition, the science and research policies should be oriented in the two main objectives:

- to assess the possibilities and needs of private and public enterprises with respect to research and technological activities.
- to choose these specific priority objectives that can describe clearly the lines of government technological action, (Korres G. 2012).

The participation of EU member states in the EU research and technological programmes can arise the opportunities for promotion and improvement of research activities, the creation of new research institutions, in order to support the innovation and the diffusion of new technologies and therefore to improve the level of economic and regional growth and social development.

REFERENCES

- Acs, J., Audretsch, B., Lehmann, E., & Licht, G. (2017). National Systems of Innovation. *The Journal of Technology Transfer*, 42(5), 997–1008. doi:10.1007/10961-016-9481-8
- Amable, B. (2000). Institutional complementarity and diversity of social systems of innovation and production. *Review of International Political Economy*, 7(4), 645–687. doi:10.1080/096922900750034572
- Bergek, A., Jacobsson, S., Carlsson, B., Lindmark, S., & Rickne, A. (2008). Analyzing the functional dynamics of technological innovation systems: A scheme of analysis. *Research Policy*, 37(3), 407–429. doi:10.1016/j.respol.2007.12.003
- Björn, J., Edquist, C., & Lundvall, B.-Å. (2003). Economic Development and the National System of Innovation Approach. In *Handbook of Innovation*. Oxford University Press.
- Carvalho, N., Carvalho, L., & Nunes, S. (2015). A methodology to measure innovation in European Union through the national innovation system. *International Journal of Innovative Research and Development*, 6(2), 159–180. doi:10.1504/IJIRD.2015.069703
- Edquist, C. (2004). Systems of innovation: Perspectives and challenges. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford Handbook of Innovation*. Oxford University Press.
- Edquist, C., & Hommen, L. (2008). *Comparing national systems of innovation in Asia and Europe: Theory and comparative framework*. In C. Edquist & L. Hommen (Eds.), *Small Country Innovation Systems: Globalization, Change and Policy in Asia and Europe*. Edward Elgar.
- European Commission. (2010). Communication from the Commission, Europe 2020, European strategy for smart, sustainable and inclusive growth. COM(2010) 2020, Brussels.

- Filippetti, A. & Archibugi, D. (2011). Innovation in Times of Crises: National Systems of Innovation, structure, and demand. *Research Policy*, 40, 179-192.
- Freeman, C. (1991). Networks of innovators: A synthesis of research issues. *Research Policy*, 20(5), 499–514. doi:10.1016/0048-7333(91)90072-X
- Freeman, C. (1995). The ‘National System of Innovation’ in historical perspective. *Cambridge Journal of Economics*, 19(1), 5–24.
- Freeman, C. (2002). Continental, national and sub-national innovation systems – Complementarity and economic growth. *Research Policy*, 31(2), 192–211. doi:10.1016/S0048-7333(01)00136-6
- Korres, G. (2012). *Handbook of Innovation Economics*. Nova Publishers.
- Korres, G., & Tsombanoglou, G. (2003). National System of Innovations in E.U., Institutional Harmonization and the effects on Sustainable Development. In Proceedings, Defining and Measuring Knowledge, New Economy Statistical Information System (Nesis Project). Luxemburg: Academic Press.
- Lundvall, B.-Å. (1988). Innovation as an Interactive Process: From User-producer Interaction to the National System of Innovations. In G. Dosi, C. Freeman, R. Nelson, G. Silverberg, & L. Soete (Eds.), *Technical Change and Economic Theory* (pp. 349–369). Pinter.
- Lundvall, B.-A. (1992). *National systems of innovation: Toward a Theory of innovation and Interactive Learning*. Anthem Press.
- Marzenna, A. (2018). Strengthening the knowledge base for innovation in the European Union. Polish Scientific Publishers PWN.
- Mytelka, L. K., & Smith, K. (2002). Policy learning and innovation theory: An interactive and co-evolving process. *Research Policy*, 31(8-9), 1467–1479. doi:10.1016/S0048-7333(02)00076-8
- Nelson, R. (1992). National Innovation Systems: A Retrospective on a Study. *Industrial and Corporate Change*, 2(2), 347–374. doi:10.1093/icc/1.2.347
- Nelson, R. R. (1993). *National Systems of Innovation: A Comparative Analysis*. Oxford University Press.
- Nunes, A., & Balsa, C. (2013). Clustering entrepreneurship aspirations: Innovation, growth and international orientation of activities. *International Journal of Innovative Research and Development*, 5(2), 165–178. doi:10.1504/IJIRD.2013.055235
- Nunes, S., Carvalho, L., & Costa, T. (2013). Cooperation for innovation: Evidence from southern European countries. *International Journal of Innovative Research and Development*, 5(2), 226–241. doi:10.1504/IJIRD.2013.055250
- OECD. (1981a). *Proposed Standard Practice for Surveys of Research and Experimental Development: “Frascati Manual 1980”, The Measurement of Scientific and Technical Activities Series*. Organisation for Economic Co-operation and Development.
- OECD. (1981b). *The measurement of scientific and technical activities: Frascati Manual 1980*. Organisation for Economic Co-operation and Development.
- OECD Organisation for Economic Co-operation and Development. (2010a). *The OCDE Innovation Strategy— Getting a Head Start on Tomorrow*. OECD Publishing.

OECD Organisation for Economic Co-operation and Development. (2010b). *OECD Science, Technology and Industry Outlook*. OECD Publishing.

ADDITIONAL READING

Korres, G. M. (1996). *Technical change and Economic Growth: an empirical evidence from European countries*. Avebury-Ashgate.

KEY TERMS AND DEFINITIONS

National Innovation Systems: Is the flow of technology and information among people, enterprises and institutions which is key to the innovative process on the national level.

Research and Development (R&D): Refers to innovative activities undertaken by corporations or governments in developing new services or products, or improving existing services or products.

Technological Policy: Concerns the public means for nurturing those capabilities and optimizing their applications in the service of national goals and interests.

Comparative Analysis of Applying Behavioral Public Policy to Telecommunication Market by International Organizations

12

Nagayuki Saito

 <https://orcid.org/0000-0003-2528-8955>

International Professional University of Technology in Tokyo, Japan

BACKGROUND

In the long history of modern society, the exercise of government regulation by the state has played an important role in the governance of its own country. Government regulation is still inevitable, such as responding to external diseconomies, dealing with disadvantages arising from an asymmetry of information and responding to moral hazards occurring in economic activities in the market.

Government regulation is authoritarian and enforcement absolute because it is regulated by the country's highest authority agency (Dye 1995). Government regulation established through such decision-making process creates a situation where law enforcement power and policy enforcement power are extremely strong. Such a policy system has allowed command and control type governance by a strict policy enforcement body of the government.

In the real society so far, government regulation by such a system has been carried out. However, recent developments in the globalization of the economy, the diversification of the values of citizens, and raising awareness of individual rights of individuals, the way of regulation centered on government regulation is shaking.

INTRODUCTION OF BEHAVIORAL PUBLIC POLICY AS A SUBSTITUTE FOR GOVERNMENTAL REGULATION

In recent years, as a new measure to compensate for the limitations of government regulations, interest in behavioral public policy has been increasing internationally. Behavior Public policy is based on the policy philosophy advocated by Thaler & Sunstein (2003). They advocate libertarian / paternalism as a concept combining contradictory social governance ideals of “libertarianism” and “paternalism.” This concept is a paternalism of relatively weak, soft, non-pushing forms and freedom of choice is not hindered; choices are not limited and do not become a heavy burden.

Thaler and Sunstein suggested that, in order to realize social governance through libertarian paternalism, “the nudge” is effective as a means of transforming people’s behavior. They argue that the nudge is every element of “the choice architecture” that predictably changes the behavior of people without forbidding people’s choices and without significantly changing economic incentives (Thaler & Sunstein 2008). It is intended that by improving the environment in which people choose, they can make wiser choices without restricting any of their options (Thaler, 2018). “The choice architecture” is an interdisciplinary

DOI: 10.4018/978-1-7998-3479-3.ch106

approach that combines and presents information to people's decision-making by providing information. In the field of cognitive psychology this information includes framing¹, anchor ring², priming³ etc., in the field of social psychology it includes inducements such as social norm⁴, bandwagon⁵ etc., and in contract theories it includes default rules⁶, opt-out schemes etc.

Oliver (2013) defines the behavioral public policy as policy measures that cause people's behavioral change to achieve the purpose of public policy. In other words, by incorporating Nudge into the public policy so far, it is possible to say that "a possible method as a way of obtaining regulatory results without setting rules, that is, a regulation other than the regulations designed to achieve the same public policy objective (Lunn 2014) "to achieve the goal of public policy.

Behavior public policy utilizing nudge is experimentally adopted in Western countries. In the UK, the Behavior Insights Team (BIT) was established in the cabinet under the direction of Prime Minister Cameron at that time, and its utilization is carried out in various policy areas. In particular, it was introduced to the policy theme, which is difficult for government intervention to individual freedom of choice, such as matters related to citizen's lifestyle and values.

In addition, in the United States, the Social and Behavioral Science Team was established in the White House, and the Presidential Decree aiming to utilize the insight of behavioral science was promulgated in September 2015. The Presidential Decree instructs us to consider whether it is possible to replace existing regulations with nudge's approach (The White House 2015).

This trend toward policy introduction of behavioral public policy is spreading to other countries as well. Such as Austria, Canada, Finland, Iceland, Ireland, New Zealand, Norway, France, Israel, Denmark and others. In addition, in the European Union which is an international organization also adapting to the public policy for member countries. Under these circumstances, the OECD is also beginning to compare practical situations to the behavioral public policy in member countries (Lunn 2014).

NECESSITY OF BEHAVIORAL POLICY TO TELECOMMUNICATION POLICY

Behavior Public policy has been mainly introduced on social welfare policies such as pension, health, medical care and public health, educational policies to raise the rate of advancement, environmental policies such as conservation of the natural environment, consumer protection policies in the market, and Telecommunication policy etc. (Lunn 2014, Thaler & Sunstein 2008, European Commission 2016, The world Bank 2015, UNCTAD 2012, The Behavioral Insights Team 2014).

As an overview of the introduction example to the telecommunication policy, in the UK, policy is taken from the viewpoint of behavioral economics in the consumer protection policy on the utilization of personal data. The initiative called "Midata" in the UK introduces regulations that force mobile communications carriers to disclose data in the event of a request for data disclosure in 2013. This regulation obliges service providers to display the main points of information in an easy-to-understand manner so that consumers can visually and easily determine their usage status (Behavioural Insights Team 2014).

Also, at the Consumer Rights Directive (CRD) of the European Union (EU), are taking consumer protection policy from the viewpoint of behavioral economics. As a case example, when buying Microsoft's operating system, it was regarded as a problem that a check button was checked in the dialog beforehand. In order to deal with this problem, CRD regulates such display. This regulation restricts such transactions that the seller side will be extremely advantageous based on the default theory that humans unconsciously choose defaults (EU 2011).

As seen in these examples, regulatory policies have been taken from the perspective of behavioral public policy in Telecommunication Market in European countries. Therefore, in this paper, will examine how behavioral public policy is adopted at international organizations which lead the public policy in international society. Furthermore, we examine the effectiveness of adopting behavioral public policy in telecommunication policy. This is because regulating the use of e-commerce and information services also touches individual values and rights. Although it is necessary to resolve policy issues, it should be avoided to take legal restrictions as much as possible. By using behavioral public policy methods, it may be possible to avoid individual values and rights restrictions.

DIRECTION OF VERIFICATION OF BEHAVIORAL PUBLIC POLICY IN TELECOMMUNICATION FIELD

Behavior Public policy leads the citizens themselves to take desirable actions as society by taking a nudge for social problems arising from people's cognitive bias without taking legal restrictive solutions (Thaler & Sunstein 2008). Because of this nature, the introduction of behavioral public policy in telecommunication policy is mainly used from the viewpoint of consumer protection.

Therefore, this paper aims to compare and verify policy stages toward introduction of behavioral public policy in consumer protection policy at each international organization. For this purpose, this study uses text analysis to examine the policy recommendations, guidelines, and reports on consumer protection policies based on behavioral insights from international organizations. Based on the analysis results, the stages at which the international agency took on consumer protection policies were classified into the following three stages:

- 1) The government recognizing the existence of consumer's cognitive bias and taking steps to remove the bias.
- 2) Verifying the social effect of the elimination of cognitive bias.
- 3) Improving the policy environment so that the government and relevant stakeholders can cooperate on the policy.

Based on this classification, aim to clarify which policy stage each international organization is leveled.

COMPARATIVE VALIDATION OF BEHAVIORAL PUBLIC POLICY IN EACH INTERNATIONAL ORGANIZATION

In this section, the main international organizations analyzed were the United Nations Conference on Trade and Development (UNCTAD), dealing with trade issues in the United Nations, the European Commission (EC), as a policy enforcement body in Europe, and the Organization for Economic Cooperation and Development (OECD), dealing with international economic issues. Specifically, publicly announced policy recommendations, guidelines, and reports of these institutions were analyzed. Based on the analysis results, the stage of the behavioral public policies in the consumer protection policy of these international organizations was compared.

Text Analysis for UNCTAD (2014) “The Benefit of Competition Policy for Consumers”

For UNCTAD, “The Benefit of Competition Policy for Consumers,” which is the UN’s consumer protection guideline (UNCTAD, 2014), was analyzed. UNCTAD points out the possibility of cognitive biases when consumers buy goods and as a guideline for UN member countries, demonstrates the need to adopt consumer protection policies based on the occurrence of cognitive bias (UNCTAD 2012, 2014).

The text analysis of this guideline revealed that 7,364 words were extracted, the number of terms type was 1,428, the average number of occurrences for the words was 3.48, and the standard deviation of the appearance frequency was 10.12. The order of the frequently appearing nouns was “consumer (208),” “competition (171),” “market (87),” “policy (76),” and “authority (51)” (see: table 1). From the numerical relationship of these frequently occurring words, it can be understood that this guideline refers to the direction of policy by the authority organization on the problem of consumer protection occurring in the course of market competition.

Table 1. Frequency of nouns in the UNCTAD report (2014)

Rank	Noun	Occurrence	Rank	Noun	Occurrence	Rank	Noun	Occurrence
1	consumer	208	8	law	28	15	benefit	21
2	competition	171	8	service	28	16	cartel	19
3	market	87	10	cost	25	16	case	19
4	policy	76	10	example	25	18	practice	18
5	authority	51	10	product	25	19	cent	17
6	price	32	13	information	24	20	choice	16
7	enforcement	30	14	benefit	21	20	decision	16

Next, in order to clarify how the elements of behavioral insight are related to this guideline, a co-occurrence network analysis was conducted. Figure 1 illustrates the co-occurrence relationship of the networks between the texts. From these co-occurrence relations, as a factor to be considered in market competition policy, consumer protection should be considered a “bias” problem that occurs in human “behavior.”

Furthermore, in order to clarify the co-occurrence relation with behavioral insight numerically, the Jaccard coefficients between each related noun and “behavioral” were calculated. The formula for calculating the Jaccard coefficient is as follows:

$$J(A, B) = |X \cap Y| / |X \cup Y|$$

The results of the analysis demonstrated that co-occurrence relationships between words related to “biases” such as “default (0.5),” “loss-aversion (0.5),” “over-optimism (0.5),” “overload (0.5),” and “time- inconsistency (0.5)” occurred in human decision-making (see Table 2).

From these results, it can be understood that consumer protection for UNCTAD (2014) is at the first stage, where the government recognizes the existence of consumer’s cognitive bias and takes steps to remove the bias.

Figure 1. Co-occurrence network of the UNCTAD report (2014)

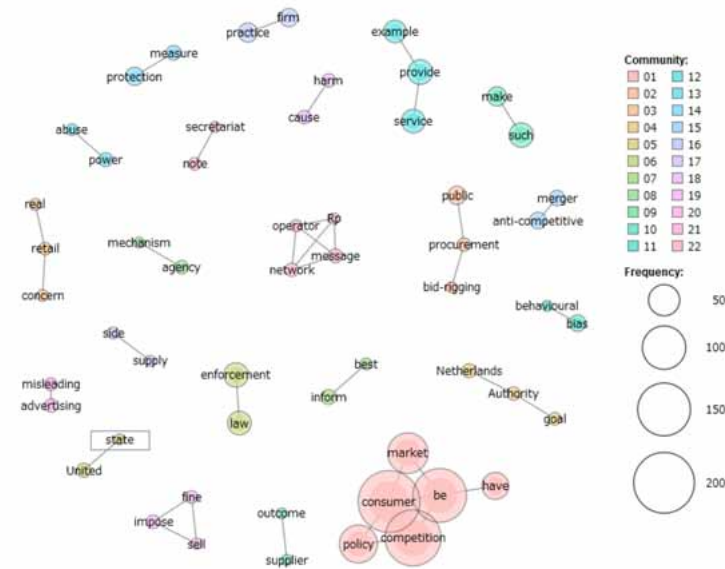


Table 2. Jaccard coefficient between “behavioral” and each word of the UNCTAD report (2014)

No.	Word	Part of speech	Unconditional	Conditional	Jaccard
1	Asset	Noun	1 (0.015)	1 (0.500)	0.5
2	Bias	Noun	1 (0.015)	1 (0.500)	0.5
3	Concentration	Noun	1 (0.015)	1 (0.500)	0.5
4	Control	Noun	1 (0.015)	1 (0.500)	0.5
5	Default	Noun	1 (0.015)	1 (0.500)	0.5
6	Loss-aversion	Noun	1 (0.015)	1 (0.500)	0.5
7	Over-optimism	Noun	1 (0.015)	1 (0.500)	0.5
8	Overload	Noun	1 (0.015)	1 (0.500)	0.5
9	Projection	Noun	1 (0.015)	1 (0.500)	0.5
10	Time-inconsistency	Noun	1 (0.015)	1 (0.500)	0.5

Text Analysis of the EU Report “Behavioural Insights Applied to Policy”

For the EU, the recommendation of the “Behavioural Insights Applied to Policy” report is subject to analysis (Lourenco et al., 2016). In this report, practical cases of behavioral public policy in each member country were collected and disseminated. Furthermore, this report makes recommendations from the viewpoint of evidence-based policies, and in addition to traditional market surveys and social surveys, demonstrates the necessity of developing evidence-based policies by using action histories and for policymakers and researchers to use this behavior data (Lourenco et al., 2016).

A text analysis of the recommendation section, “4. Lessons learnt and recommendations,” in this report revealed that 1,023 words were extracted, the number of terms type was 401 words, the average number of occurrences was 1.72, and the standard deviation of the appearance frequency was 2.36. The

order of the frequently appearing nouns was “policy (20),” “intervention (7),” “country (5),” “initiative (5),” and “instance (5)” (See Table 3).

Table 3. Frequency of nouns in the Lourenco et al. report (2016)

Rank	Noun	Occurrences	Rank	Noun	Occurrences	Rank	Noun	Occurrences
1	policy	20	8	awareness	3	8	policy-maker	3
2	intervention	7	8	community	3	8	policy-making	3
3	country	5	8	enforcement	3	8	process	3
3	initiative	5	8	evaluation	3	8	research	3
3	instance	5	8	evidence	3	8	researcher	3
6	effectiveness	4	8	implementation	3	8	result	3
6	impact	4	8	issue	3	8	section	3

Next, in order to clarify how the elements of behavioral insight are related, a co-occurrence network analysis was conducted (see Figure 2). From this co-occurrence relation, the assertion of this document that it is important to adduce evidence to evaluate the effect of applying policy using behavioral insight was demonstrated.

Furthermore, in order to clarify the co-occurrence relationship with behavioral insight numerically, the Jaccard coefficient between each relation noun and “behavioral” was calculated. From this, the words having a relatively high Jaccard coefficient were “policy (0.46),” “initiative (0.33),” “awareness (0.22),” “section (0.22),” and “way (0.22).” Based on the results of this analysis, it can be understood that this recommendation requires the initiative and utilization of public policy using behavioral insight (see Table 4).

Figure 2. Co-occurrence network of the Lourenco et al. report (2016)

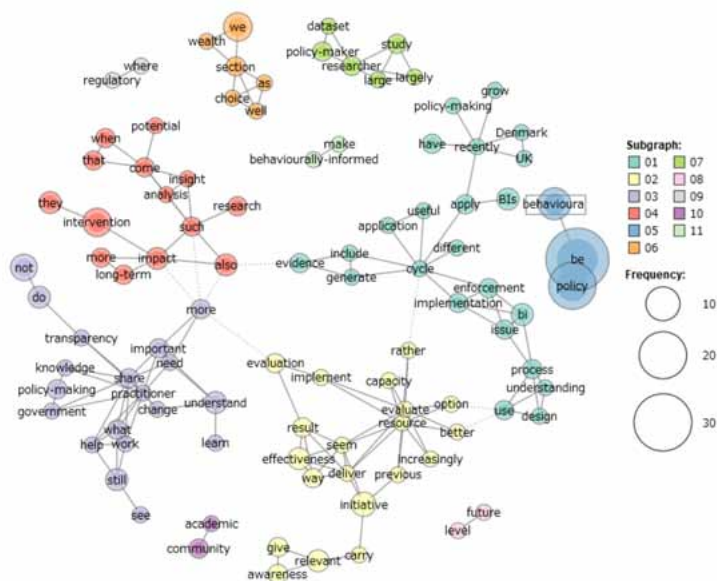


Table 4. Jaccard coefficient between “behavioral” and each word of the Lourenco et al. report (2016)

No.	Word	Part Of Speech	Unconditional	Conditional	Jaccard
1	policy	Noun	11 (0.579)	6 (0.750)	0.4615
2	BIs	Proper Noun	4 (0.211)	3 (0.375)	0.3333
3	initiative	Noun	4 (0.211)	3 (0.375)	0.3333
4	awareness	Noun	3 (0.158)	2 (0.250)	0.2222
5	section	Noun	3 (0.158)	2 (0.250)	0.2222
6	way	Noun	3 (0.158)	2 (0.250)	0.2222
7	account	Noun	1 (0.053)	1 (0.125)	0.125
8	activity	Noun	1 (0.053)	1 (0.125)	0.125
9	appetite	Noun	1 (0.053)	1 (0.125)	0.125
10	asset	Noun	1 (0.053)	1 (0.125)	0.125

From this, it can be determined that the EU Consumer protection as of 2016 is in the second stage of verifying the social effect of the elimination of cognitive bias.

Text Analysis for OECD (2016) “Consumer Protection in E-commerce OECD Recommendation”

In 2016, the OECD published the “Consumer Protection in E-commerce OECD Recommendation” as a consumer protection recommendation for e-commerce. Based on the current situation that behavioral information is utilized in e-commerce, the following recommendation was made in this report. As a general principle for consumer protection, in order to develop “transparent and effective protection,” and to protect children, vulnerable people, and disadvantaged consumers, depending on the special circumstances of e-commerce, stakeholders and governments need to gain insight from information and behavioral economics. For that reason, from the viewpoint of behavioral insight, it is necessary to obtain behavioral information, and in order to achieve the purpose of these recommendations, as the enforcement principle, governments should impose “empirical research based on insight gained from behavioral economics” in cooperation with stakeholders (OECD 2016).

A text analysis of this recommendation revealed that there were 4,732 extracted words, the number of terms type was 1,003 words, the average number of occurrences was 3.06, and the standard deviation

Table 5. Frequency of nouns in the OECD report (2016)

Rank	Noun	Occurrences	Rank	Noun	Occurrences	Rank	Noun	Occurrences
1	consumer	123	8	payment	21	14	stakeholder	14
2	business	72	9	service	20	16	resolution	13
3	e-commerce	45	10	government	18	17	authority	12
4	information	41	10	mechanism	18	18	risk	11
5	transaction	36	12	dispute	16	19	access	10
6	protection	31	12	practice	16	19	goods	10
7	recommendation	30	14	enforcement	14	19	redress	10

tion of the number of occurrences was 6.66. The order of frequently appearing nouns was “consumer (123),” “business (72),” “e-commerce (45),” “information (41),” and “transaction (36)” (See Table 5).

Next, in order to clarify how the behavioral insight elements are related in this report, a co-occurrence network analysis was conducted (See: Figure. 3). From the co-occurrence relationship illustrated in Figure 3, from the viewpoint of consumer protection, to solve technical problems in e-commerce, knowledge of behavioral economics to improve the environment in which the government and stakeholders collaborate must be utilized.

Furthermore, the Jaccard coefficient between each related noun and “behavioral” was calculated in order to clarify the co-occurrence relationship with behavioral insight quantitatively. As a result of this analysis, a strong relationship was demonstrated in nouns such as “economics (1.0),” “insight (1.0),” “evidence (0.75),” “base (0.5),” and “research (0.5).” From this result, the relationship of the nouns

Figure 3. Co-occurrence network of the OECD report (2016)

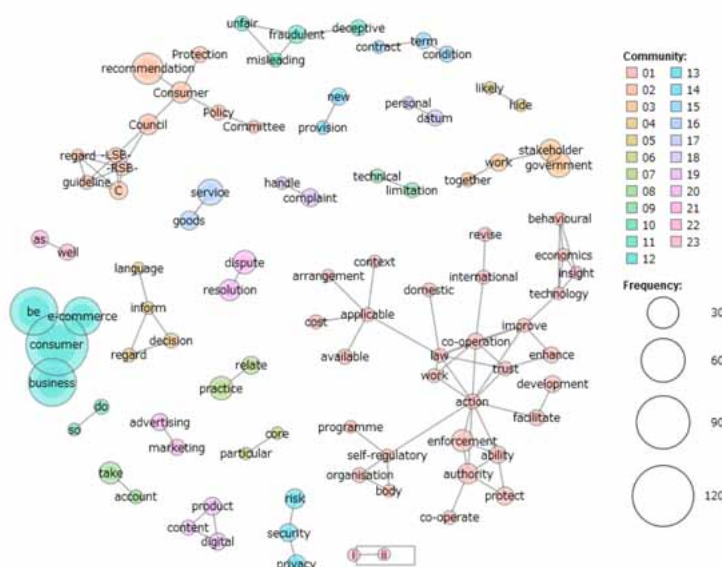


Table 4. Jaccard coefficient between “behavioral” and each word of the OECD report (2016)

No.	Word	Part Of Speech	Unconditional	Conditional	Jaccard
1	economics	Noun	4 (0.038)	4 (1.000)	1
2	insight	Noun	4 (0.038)	4 (1.000)	1
3	evidence	Noun	3 (0.029)	3 (0.750)	0.75
4	base	Noun	2 (0.019)	2 (0.500)	0.5
5	research	Noun	2 (0.019)	2 (0.500)	0.5
6	technology	Noun	3 (0.029)	2 (0.500)	0.4
7	policy	Noun	8 (0.077)	3 (0.750)	0.3333
8	addition	Noun	1 (0.010)	1 (0.250)	0.25
9	advice	Noun	1 (0.010)	1 (0.250)	0.25
10	analysis	Noun	1 (0.010)	1 (0.250)	0.25

indicating the necessity of consumer protection policy based on a user's behavioral history was demonstrated (see Table 6).

From these results, a data analysis based on the IoT environment, which analyzes the purchase/behavior history obtained from account histories, is recommended for consumer protection policies in the OECD. In addition, this is recommended for policymakers and researchers and also to urge cooperation between supplier companies utilizing the IoT.

Therefore, it can be said that the OECD's policies are in the third stage, improving the policy environment so that the government and relevant stakeholders can cooperate with the policy.

SOLUTIONS AND RECOMMENDATIONS

From these results, it can be understood that: UNCTAD (2014) consumer protection is at 1) the stage where the government recognizes the existence of the consumer's cognitive bias and takes steps to remove the bias, EU consumer protection as of 2016 is at 2) the stage of verifying the social effect of the elimination of cognitive bias, and the OECD (2016) is at 3) the stage of improving the policy environment so that the government and relevant stakeholders can cooperate with the policy (See Table 7).

Table 7. Comparison of consumer protection policies using behavioral public policy by each international organization

	1) Recognition of consumer's cognitive bias	2) Verification of the effect of behavioral public policy	3) Improvement of collaborative infrastructure with stakeholders
UNCTAD	Mentioned the necessity of developing a consumer protection policy based on the occurrence of cognitive bias.	-	-
EU	Collecting and sharing practical examples of behavioral public policy in member countries.	Verification of the social effect of the elimination of cognitive bias.	-
OECD	Understand the current situation that behavior information is utilized in e-commerce.	Developing optimal consumer protection by analyzing consumer's behavioral histories.	Cooperate with stakeholders to improve e-commerce policy based on behavioral evidence.

Based on these results, the recommendations for Japan follow. In Japan, the Ministry of the Environment launched the Behavioral Sciences Team (BEST) in 2017, and the Cabinet Office, the Ministry of Health, Labor and Welfare, the Ministry of Internal Affairs and Communications, and the Ministry of Land, Infrastructure, and Transport, among others, are cooperating with BEST and considering public policy based on behavioral insights in Japan. In 2018, BEST was at the stage of reviewing the public policy behavioral insights of European countries and the U.S. and accumulating domestic cases to demonstrate the effect of these insights in Japan.

Therefore, it can be said that Japan is transitioning from stage 1) to stage 2) of the public policy based on behavioral insights. In order to advance to the third stage 3) and improve the environment for government and stakeholders to cooperate on policy, it seems necessary to strengthen the cooperation of industry, government, and academia through BEST.

FUTURE RESEARCH DIRECTIONS

From the results of the above analysis, the stages of application of behavioral insights to the telecommunication policy in each international organization emerged clearly. In order to further evolve this research, it will be effective to structurally analyze cases of information communication policies in each country where behavioral insights are applied.

For instance, OECD's Public Governance and Territorial Development Directorate Regulatory Policy Division conducted surveys in member countries to comprehensively examine the implementation and effectiveness of public policies that applied behavioral insights in the international community. According to their report, 159 policy cases were reported using examples of telecommunication policies in countries including the U.K., Colombia, and the Netherlands (OED, 2017b).

Quantitative and qualitative analysis on the textual data of this report will help characterize and clarify various aspects of the telecommunication policy using behavioral insights. Such analysis will reveal the kinds of behavioral insights that are applied to telecommunication policy issues, why it is effective to apply behavioral insights, human habitual problems that may occur, and possible solutions to these problems by the application of behavioral insights.

In addition, the OECD has presented some forms of cognitive biases that must be taken into consideration when implementing consumer protection policies in e-commerce (OECD, 2017a) along with some suggested "de-biases" to counter the occurrence of biases (OECD, 2018). In order to develop this research, the next step mentions the effectiveness of the consumer protection policy to prevent the occurrence of cognitive bias that might arise from the characteristics of e-commerce and help consumers make appropriate choices online.

CONCLUSION

This study examined behavioral public policy as a new policy approach and alternative to how conventional government regulation is incorporated into public policy in Western countries. The U.K. and the U.S. are undertaking advanced measures in developing behavioral public policy at the national level, and following this, European countries are also actively taking measures of their own to develop public policy. Furthermore, this study confirmed that international organizations are recommending the introduction of international behavioral public policy for formulating policy and guidelines.

Based on these international policy trends, from previous literature reviews, this study clarified that behavioral public policy is a policy approach for international organizations. Furthermore, this paper conducted a comparative study of the literature survey results of the stage of the introduction of behavioral public policy at each international organization.

Using these results, previous evidence-based policies were measured by using data from social surveys and market surveys. By introducing the behavioral public policy approach, it will be possible to judge what type of nudge is effective as policy. Based on behavioral evidence, it is also possible to judge whether consumer behavior had changed in the appropriate direction or whether it was possible to protect consumers by using nudges.

In order to use this information on communication policy, cooperation between governments and relevant industries is indispensable. Thus, this study demonstrated a new policy direction based on evidence by considering the current state of international behavior public policy, and in doing so contributes to information and communication policy.

REFERENCES

- Dye, T. R. (1995). *Understanding Public Policy* (7th ed.). Prentice-Hall.
- European Union. (2011). Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7, EC of the European Parliament and of the Council.
- Lourenço, J. S., Ciriolo, E., Almeida, S. R., & Troussard, X. (2016). *Behavioural Insights Applied to Policy: European Report 2016*. European Union.
- Lunn, P. (2014). *Regulatory Policy and Behavioural Economics*. OECD Publishing. doi:10.1787/9789264207851-en
- Ministry of Environment. (n.d.). *Behavioral Sciences Team (BEST)*. <http://www.env.go.jp/earth/best.html>
- OECD. (2016). *Consumer Protection in E-commerce OECD Recommendation*. OECD Publishing.
- OECD. (2017a). *Use of Behavioral Insights in Consumer Policy*. OECD Publishing.
- OECD. (2017b). *Behavioral Insights and Public Policy - Lessons from Around the World*. OECD Publishing.
- OECD. (2018). *Improving online disclosures with behavioral insights*. OECD Publishing.
- Oliver, A. (2013). *Behavioural public policy*. Cambridge University Press. doi:10.1017/CBO9781107337190
- Social and Behavioral Science Team. (n.d.). <https://sbst.gov/>
- Thaler, R. H. (2018). Nudge, not sludge. *Science*, 361(6401), 431. doi:10.1126/science.aau9241 PMID:30072515
- Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism Is Not an Oxymoron. *The University of Chicago Law Review. University of Chicago. Law School*, 70(4), 1159–1202. doi:10.2307/1600573
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press.
- The Behavioural Insights Team. (2014). *EAST: Four simple ways to apply behavioural insights*. The Behavioural Insights Team. <http://www.behaviouralinsights.co.uk/>
- The White House. (2015). *Executive Order -- Using Behavioral Science Insights to Better Serve the American People*. <https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american>
- The World Bank. (2015). *Mind, Society, and Behavior*. International Bank for Reconstruction and Development / The World Bank.
- UNCTAD. (2012). *Consumer protection and competition policy*. United Nations.
- UNCTAD. (2014). *The benefit of competition policy for consumers*. United Nations.

ADDITIONAL READINGS

Angner, E., & Loewenstein, G. (2012). Behavioral Economics. In U. Mäki (Ed.), *Springer* (pp. 641–689). Handbook of the Philosophy of Science.

Heukelom, F. (2015). *Behavioral Economics: A History*. Cambridge, UK. Cambridge University Press.

Nagatsu, M. (2015). History of Behavioral Economics. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., Vol. 2, pp. 443–449). Elsevier. doi:10.1016/B978-0-08-097086-8.03053-1

OECD. (2016). *OECD Recommendation of the Council on Consumer Protection in E-Commerce*. OECD Publishing.

OECD. (2016). *Protecting Consumers through Behavioural Insights- Regulating the Communications Market in Colombia*. OECD Publishing.

The Behavioural Insights Team. (2019). The Behavioural Insights Team Annual Update Report 2017-18, The Behavioural Insights Team, London, <https://www.bi.team/wp-content/uploads/2019/01/Annual-update-report-BIT-2017-2018.pdf>.

KEY TERMS AND DEFINITIONS

Behavioral Economics: Behavioral economics is economic studies that do not assume homo-economics and conduct analysis according to the reality of human psychology and emotions.

Behavioral Insights: Behavioral insights are insights gained from theories on behavioral science and social sciences such as decision theory, psychology, cognitive science, brain science, organizational behavior, and group behavior, and so on.

Behavioral Public Policy: Behavioral public policy is the development of public policy by applying behavioral insights. There are interventions aimed at transforming people's behavior into desirable behavior by using nudges and correcting cognitive bias.

Choice Architecture: Choice architecture is an interdisciplinary approach that combines and presents information on people's decision-making by providing information.

Cognitive Biases: Cognitive biases is a theory of cognitive psychology and social psychology that posits that when evaluating a certain object, an idea can be distorted in a direction according to a person's interests and hopes, and the evaluation is distorted by the features that are noticeable to the object.

Evidence-Based Policy: Evidence-based policy is a policy approach for policy planning based on numerical evidence such as statistical data and is widely adopted in various fields such as social policy, educational policy, and welfare policy.

Nudge: A nudge is a measure advocated by Thaler and Sunstein (2008) that transforms the behavior of people in an appropriate direction for society. This encourages behavioral changes using all the elements of choice architecture that do not prohibit people's choices and does not change economic incentives significantly, and so changes the behavior of people in a predictable manner.

ENDNOTES

12

- ¹ Framing: For example, when the probability of unsuccessful surgery is presented as 10%, or when the probability of successful surgery is presented as 90%, people's decisions are distorted depending on how the information is presented.
- ² Anchor ring: A human behavioral tendency that chooses close to that numerical value by presenting a certain value as an anchor.
- ³ Priming: Information given ahead of time, such as advertisements, that will affect people's decision-making and preferences.
- ⁴ Social norms: People's behaviors and decisions that are influenced by the values and expectations of the particular society or group to which they belong.
- ⁵ Band wagon: A human behavioral tendency to be influenced to make the same choice that many other people have chosen.
- ⁶ Default rules: A human behavioral tendency to select a choice by being presented with one option as the default.

A Good American President

James George

Archway Publishing, USA

Abdullah Murrar

 <https://orcid.org/0000-0002-4445-8326>

Indiana University of Pennsylvania, USA

Pankaj Chaudhary

Indiana University of Pennsylvania, USA

James Allen Rodger

Indiana University of Pennsylvania, USA

INTRODUCTION

To what extent does the content of resumes indicate the future performance of candidates for any elected position to which they are aspiring incumbency? Using the model of business firms that may very well be “best practices,” it is suggested that voters evaluate candidates for elected office in the same manner that they might be evaluated by their employers for prospective jobs (George & Rodger, 2017).

In essence, every US citizen has an opportunity, if not an obligation, to act in the role of a hiring employer when it comes to selecting and electing candidates for public office. It cannot be presumed that every citizen in the voting public knows how to do this.

To evaluate resumes of candidates for the job of president of the US, one must have a job description and must have a list of required and desired qualifications. As it stands today, there is no job description and there is no list of qualifications beyond a skimpy set of legal requirements as specified in the US Constitution.

Voters need information to make educated decisions, and it is journalists’ job to give it to them (Andra Brichacek, 2019). Through today’s news, voters can get the basic facts on the candidates through a quick internet search, many news agencies have differentiated themselves by shifting from straight news to context and analysis as well as adopting the social media. According to Pew Research Center, about 68% of American adults get news on social media even though many have concerns about its accuracy (Matsa & Shearer, 2018). Social media gives users more direct access to candidates than ever before. Thus, the role of the third party organizations or the fourth branch of government is not only quick bringing the information to the voters, but also providing reliable information about the candidates.

The likelihood of selecting the person who has coherent skill, wide knowledge, strong experience, high proficiency, good judgement, viewing wisdom to perform in such a responsible and comprehensive capacity is the most important in this process and that can deliver a good president. Therefore, it is suggested that voters would be able to require candidates to produce certifiable evidence that they have the required qualifications and the adequate competences, because the president job surely demands it. Voters accordingly expect the outcomes based on the tasks and activities the president performs according to the skills and competences.

DOI: 10.4018/978-1-7998-3479-3.ch107

So far, voters need criteria to evaluate the candidates, thus they can judge whether he or she is a good candidate or not indeed. The authors of this chapter have researched the subject of evaluating Presidential Candidates and incumbent performance as part of an on-going series of books and papers that are concurrent with their having engaged Federal Government clients. The idea in this chapter is to connect the dots among the criteria and considerations that help citizen voters elect superior candidates to public office. While the example is the Office of President, the process applies to all elected officials.

CHAPTER OBJECTIVES

Past actions and competence are the politician's records that best predict how they might govern in the future. Uncovering these records are crucial to vet candidates and choose the best one. The purpose of this chapter is to:

1. Demonstrate how to define what makes a Presidential candidate or incumbent a good one.
2. Describe the necessity for political parties and other third-party organizations to help voters vet candidates.
3. Discuss the nature of competence and character in evaluating prospective candidates and incumbents.

BACKGROUND

This paper addresses the meaning of a “good American President,” as an instance of word-sense disambiguation (WSD). What is the meaning of “good” in this use?

Defining the Context

The authors argue that the highest order purpose of government is striving to create an ideal environment in which citizens and their commercial businesses can thrive and optimize in a global economy. In a democratic republic such as the United States, governments collaborate with private corporations to maximize the return on national resources. The aim is to develop policies that optimize return on the resources of the nation.

Governments are service-oriented enterprises that attend citizens needs as prescribed by them through levels of government systems and representation. Governments are funded by tax revenues and other fees as legislated. Legislators work with executives to establish plans, programs, services, budgets, and enabling funding (National Priorities, 2016). Central governments and even local governments are expected to balance their budget outcome and to be financially sustainable (Bolívar et, al, 2015).

Departments and Agencies comprise the Federal government as bureaucratic entities created by acts of legislation. The federal government also charters a host of corporations as independent entities. States governments also charter corporations. Most local, county, and municipal governments are public, non-profit corporations.

Ecosystem services are conditions and processes through which natural ecosystems and the species that make them up sustain and fulfil the human life. Thus provide such things like agricultural produce, timber, and aquatic organisms such as fishes and crabs (Daily, Gretchen C, 1997). Economists and ecologists have been working together for several decades to find ways to explain the dependence of humans

on the natural environment, and refine language on how ecological and economic processes interconnect for the benefit of society (Steven Cork et, al, 2007). Thus, economic “National resources” which include natural resources, raw materials, land, and capital are primary inputs to a myriad of public and private processes. There are multiple sources for determining the inventory of national resources.

Processes are all of the systems and activities that constitute the work required to produce desirable and expected outcomes. They are a product of legislation and the work of the Federal government organizations.

In the American political system, the US Constitution is the overarching control, a network of laws and regulations that guide and constrain the processes, including private sector business rules. Where processes define “how” work is done, people and technology perform the actual work (U.S. History.org, 2019) .

The government systems are staffed by people in government organizations and contractors to do the work as legislated by Congress. They are “mechanisms” that transform inputs into value-adding outcomes.

The Executive branch manages government services as a co-equal partner in the process.

The court system administers justice and resolves disputes.

Citizens and corporations staff private sector processes and systems to produce products and services offered, delivered, sold, and bought through the network of market-driven commerce. People and commercial organizations perform work, where compensation is based on individual’s value-adding contributions.

A nation’s gross domestic product is a way to quantify economic productivity. Only commercial corporations and entrepreneurs generate products and services sold in the global marketplace, which create a nation’s wealth. The Gross Domestic Product (GDP) is the current measurement method (U.S. Bureau of Economic Analysis, 2019).

Alternatively, Genuine Progress Indicator (GPI) is as a better way to measure performance with respect for economic, social, and environmental responsibility (Tejvan Pettinger, 2011). In their book *Regenerating America with Sustainable Economics*, (George & Rodger, 2017) argue that capitalism is unsustainable. Suggested is a new model needed from which to engineer sustainable economies, and the authors call it the Optimized Sustainment Model (OSM).

Outcomes

“Outcomes” include all of those associated results from government services delivery and all of those produced by private corporations and individuals.

Private Sector Outcomes

- Profit (Capital)
- Products
- Services
- Customer Satisfaction
- Assets (Intellectual property)
- Expenses (Cost of doing business)
- Physical property
- Jobs/Employment
- Debt

- Interest
- Cash flow

Government (Public Sector) Outcomes

American citizens have every right to expect the following outcomes listed in relative priority. The percentage following the result suggests as what most citizens expect.

1. Human rights as defined in the Universal Declaration of Human Rights (Hilton et.al, 2007) to which the USA is a signatory supported 100%
2. The US Constitution and The Bill of Rights guaranteed 100%
3. The Rule of Law supported 100%
4. Nation secured economically - 100%
5. Nation has a sustainable economy - 100%
6. Clean air and water - 100%
7. Nation secured militarily - 100%
8. The best education in the world - #1
9. Homeland defended from foreign and domestic terrorists - 100%
10. Employment opportunities for all who can work - 100%
11. Upward mobility for all who are prepared commensurate with ability - 100%
12. All persons provided minimal sustinment as a baseline to start - 100%
13. Social security assured for all persons - 100%
14. The freest nation in the world - #1
15. The lowest crime rate in the world - #1 in the least crime per capita

Observe that “human rights” is listed first, and that is because it is inalienable to every citizen in the world independent from the form of government or state in which they live.

In common with the Private Sector Outcomes, these are the Government's:

- Revenue from Taxes and Fees
- Products
- Services
- Customer Satisfaction
- Assets (Intellectual property)
- Expenses (Cost of service delivery)
- Physical property
- Jobs/Employment
- Debt
- Interest
- Cash flow
- Safety and Security
- Health
- Infrastructure

Let's discuss the outcomes and their relative priority. America, the state, is defined by the *US Constitution* and the *Bill of Rights*. However, those elements of the American Political System are subordinated to the larger principle of human rights for every citizen on the planet. That is why this list begins with "The Universal Declaration," and that is a teaching point for many citizens because equality for all starts with those propositions and instruments. The rule of law, the restriction of the arbitrary exercise of power by subordinating it to well-defined and established laws, is essential to all the rest.

A state must be secured economically, or else it is unviable and unable to care for citizens as they expect in a democratic republic. A country must generate sufficient tax revenue to support its obligations. It must not spend more than it can afford, as reflected in the debt and deficit.

It is essential that citizens agree with the purpose of government. The primary purpose of government is to create an environment in which individuals and their corporations can thrive in the strong economy. The aim of government is to optimize return on the nation's resources in partnership with citizens and their public and private corporations. To what extent do citizens agree with that proposition?

The rule of law applies internally and externally among trading partners and allies. Without that protection, there is no financial integrity and security.

Expectations about the economy have evolved into the concept of sustainable economics as accomplished by balancing economic, social, and environmental responsibilities. Having a sustainable economy is dependent on developing and fielding renewable energy systems. Renewable energy is a national requirement to be accomplished in public and private partnership as part of the national infrastructure.

Humanity cannot survive and thrive without clean air and water. While there is a debate about climate change and environmental science, undeniably, people have spoiled the air and water to the detriment of their health.

A nation-state must secure its integrity with military power as long as a rogue, and hostile nations exist.

Having an education system that produces superior students and citizens that compete in the global economy is essential as is providing ample employment opportunity with corresponding upward mobility.

Impoverishment must be illegal. That everyone is ensured minimal sustainment as a baseline to start.

Necessities

The readers should consider these points in a very personal manner. What do people need as a matter of necessity at various stages in their life cycles?

In the beginning:

- All citizens have affordable healthcare for their lifetime
- All citizens have affordable housing, food, and access to essential transportation
- Parents have work to perform to earn a living and to attend their obligations
- Community has police and fire protection to make the home safe
- Community has excellent schools that are accessible to young citizens
- Geniuses are provided opportunities to learn at a rate commensurate with their abilities based on their performance
- All children are encouraged to perform to their optimum ability as a matter of personal responsibility and self-reliance
- Opportunities are planned for scaling to the demographics of the emerging population

Upon high school graduation:

- HS graduates have access to college to prepare for professions and vocations
- All graduates, HS and college have access to work to earn a sustainable living
- All HS graduates have access to low interest or no interest education and training loans

Upon college and training graduation or completion:

- Graduates have access to job opportunities commensurate with their abilities
- Graduates who seek to become entrepreneurs have access to capital resources commensurate with their plans and preparation
- Graduates may expect government and private enterprise to produce sufficient upward mobility to support their advancing preparation and development

As working citizens in professions, occupations, vocations, and jobs:

- Citizens may expect the opportunity to save and invest a portion of their earnings to enhance their well-being
- Citizens may expect programs from both the private sector and government to support planned retirement
- All citizens may expect to live a good life in the absence of poverty providing that they commit to performing to the best of their ability

To what extent do citizens agree that these are necessities? Is a referendum needed?

What can citizens expect from their earnest labor and for being responsible citizens? Needed is a system that ensures exceptional reward and incentives for exceptional performance. Needed is a system that contains a balance of consequences such that poor performance is discouraged, or punished. That system needs to be shared throughout society in both private and government enterprise.

Should that system be a free-for-all or a natural accidental occurrence such as by invisible hands, or should it be more deterministic by a visible government that operates by the hands of *We the People*?

Being a responsible and self-reliant citizen who exhibits social and environmental responsibility is a sustainment position in a healthy economy.

When citizens feel responsible, they are willing to improve their country continuously. One way of development is by electing qualified candidates who meet the aspirations of citizens by implementing the good system they desire. Of course, a good president essentially contributes to develop a good system.

PRESIDENT'S JOB MODEL

To determine what makes a “good president,” one must know the job tasks for the position of President. The authors applied a job modelling technique to define the presidential tasks. The concept of modelling jobs was derived from performance management science that was developed by AT&T in the 1970s in concert with numerous consultants and academics for the purpose of engineering precise performance from the combination of people and technology. That foundation served for continuous development with the advancement of computer technology and understanding about the management of work. Accomplishing the defining of the presidential tasks required referencing the US Constitution and applying knowledge from familiarity with the US Federal Government through consulting work.

Table 1. Presidential Tasks Performance

1. Planning, staffing, organizing, and scheduling Presidential work and government functions.
1.1: Recruit and staff the cabinet and department and agency appointments.
1.2: Conduct cabinet meetings to develop and implement strategies and policies for accomplishing the nation's workload and issues.
1.3: Define the nation's outcomes and priorities for each major department and agency to produce the nation's strategic plan.
2. Develop the President's management agenda and budget and reconcile with Congress.
2.1: Assume responsibility for the legacy agenda and make adjustments to align with the President's management agenda or equivalent.
2.2: Work with cabinet heads to develop performance plans and schedules.
2.3: Work with Congressional leadership and collaborate to implement the nation's strategy, plans, and programs.
3. Initiate and approve legislation
3.1: Propose bills to Congress
3.2: Consult and advise Congress
3.3: Collaborate with industry and business leaders in the development of policies and regulations of all kinds
3.4: Approve or veto legislation
3.5: Request a declaration for war from Congress
4. Implement plans and manage on-going operations
4.1: Review and evaluate programs that include new acquisitions and on-going operations
4.2: Evaluate programs including legacy processes and engineer new ones for accomplishing required and promised outcomes
4.3: Continuously improve
5. Report progress and discuss issues with the American public to keep them informed
6. Meet with heads of state and participate in international meetings and conferences for heads of state
6.1: Participate in international economic conferences
6.2: Respond to international crises and requests for assistance
6.3: Promote democratic process and reforms

Performance analysis and modelling techniques were applied to identify the six primary tasks performed by Presidents of the United States.

The Concept of Operations (CONOPS) of the Work of the President

Generally, the CONOP is a matter of document describing the characteristics of a proposed system from the viewpoint of an individual who will use that system. A CONOP helps an organization in language what is required and what should be built for an information system. Thus it addresses the system's function, the users of the system and the system's purpose (Techopedia, 2019). The concept of operations (CONOPS) of the American Political System begins with the US Constitution that delineates the national frame of government to act on a national level.

Statement of the Goals and Objectives of the System:

The President is the Chief Executive of the Federal government enterprise and is the Commander in Chief presiding over all Departments and Agencies as well as the White House Staff.

Primary threads of responsibility include:

1. Staffing the Executive positions of all Departments and Agencies subject to Senate approval.
2. Managing Departments and Agencies to ensure the delivery of government services as prescribed by laws and regulations.
3. Planning and budgeting, establishing priorities reflected in the President's management agenda, in concert with Congress.
4. Making and changing policies, including issuing Executive Orders.
5. Strategizing, predicting, forecasting, decision making, and sense making

Managerial skills in a contemporary paradigm include the following:

- Collaborating and facilitating (aka organizing)
- Acquisition, logistics, procurement
- Financing
- Safety and security
- Strategies, tactics, policies, and constraints affecting the system
- Organizations, activities, and interactions among participants and stakeholders
- Clear statement of responsibilities and authorities delegated
- Specific operational processes for fielding the system
- Processes for initiating, developing, maintaining, and retiring the system (laws and regulations)

Candidate and Incumbent Resume and Certifications

Citizen-voters, political parties, and the media have the responsibility to vet candidates and incumbents based on their past performance and experience as aligned with the job model and statutory requirements. The process is a matter of verifying the resume and claims made, as would anyone seeking employment. For anyone wanting a job in government service, there are standards including those for Senior Executive Service, for instance. Why should citizens expect anything less for the President or any other elected official?

Therefore, the authors derived suggested skill, knowledge, and experience requirements and presented them in an evaluation matrix. Included also are considerations for a candidate's character, mental, and physical health where board-certified professionals are sought for assessment and certification. Consideration is given intelligence (IQ) in a similar manner.

Financial and legal integrity are sought just as for any federal employee requiring a security clearance. Why would not a candidate for elected office, including the President, not be subject to the same scrutiny as other government employees?

Table 3 in appendix proposes the Presidential Candidate Evaluation. The matrix contains the evaluation criteria, the value and the score of each criterion. The Qualifications Required by Law: As directed by the Constitution such as age and nationality of the President. This matrix also incorporates the President's academic qualifications, work experience, work nature, personality, knowledge with government, military background and other parameters.

Citizen-Voters make judgments based on reviewing candidate resumes as vetted and certified, and based on other information such as a President's Manifesto and alignment with Party Platforms. They produce their weighted score for comparison with other candidates under consideration.

Intelligence Quotient (IQ) is omitted from the criteria, though a board-certified declaration might be helpful. Additional research is necessary to support this topic.

DISAMBIGUATING THE MEANING OF A “GOOD AMERICAN PRESIDENT”

Issues, Controversies, Problems

Like many voters in other democratic societies, American voters are ill-prepared for their responsibility to elect officials, much less vote on issues and levies. The average American voter operates with a tenth-grade education, for instance. At no time are students instructed about how to evaluate elected officials in a manner that is consistent with the process employed by The Office of Personnel Management (OPM). As an independent agency, OPM manages the civil service of the federal government, coordinates recruiting of new government employees, and manages their health insurance and retirement benefits programs. According to the Senior Executive Policy of (Office of Personnel Management OPM, 2019), the OPM provides the day-to-day oversight and assistance to agencies as they develop, select, and manage their Federal executives. Also Members of the SES serve in the key positions just below the top Presidential appointees and they operate and oversee nearly every government activity.

“The Senior Executive Service (SES) lead America’s workforce. As the keystone of the Civil Service Reform Act of 1978, the SES was established to “...ensure that the executive management of the Government of the United States is responsive to the needs, policies, and goals of the Nation and otherwise is of the highest quality.” These leaders possess well-honed executive skills and share a broad perspective on government and a public service commitment that is grounded in the Constitution. (Office of Personnel Management OPM, 2019).

Now and never has there been a Job Application for President of the United States. That’s a glaring omission.

Issue #1: Standards are ill-defined and incomplete for evaluating elected officials including the Office of President of the United States.

Issue #2: Standards such as suggested in this chapter should result from collaboration and consensus among political parties and with implications for Constitutional amendments governing the political process.

Issue #3: Staffing the Executive branch demands timely and complete action to ensure that all Departments and Agencies have essential leadership. The process is made slow by incumbent executives being ill-prepared to staff the government and by the security Clearance process requiring time to complete.

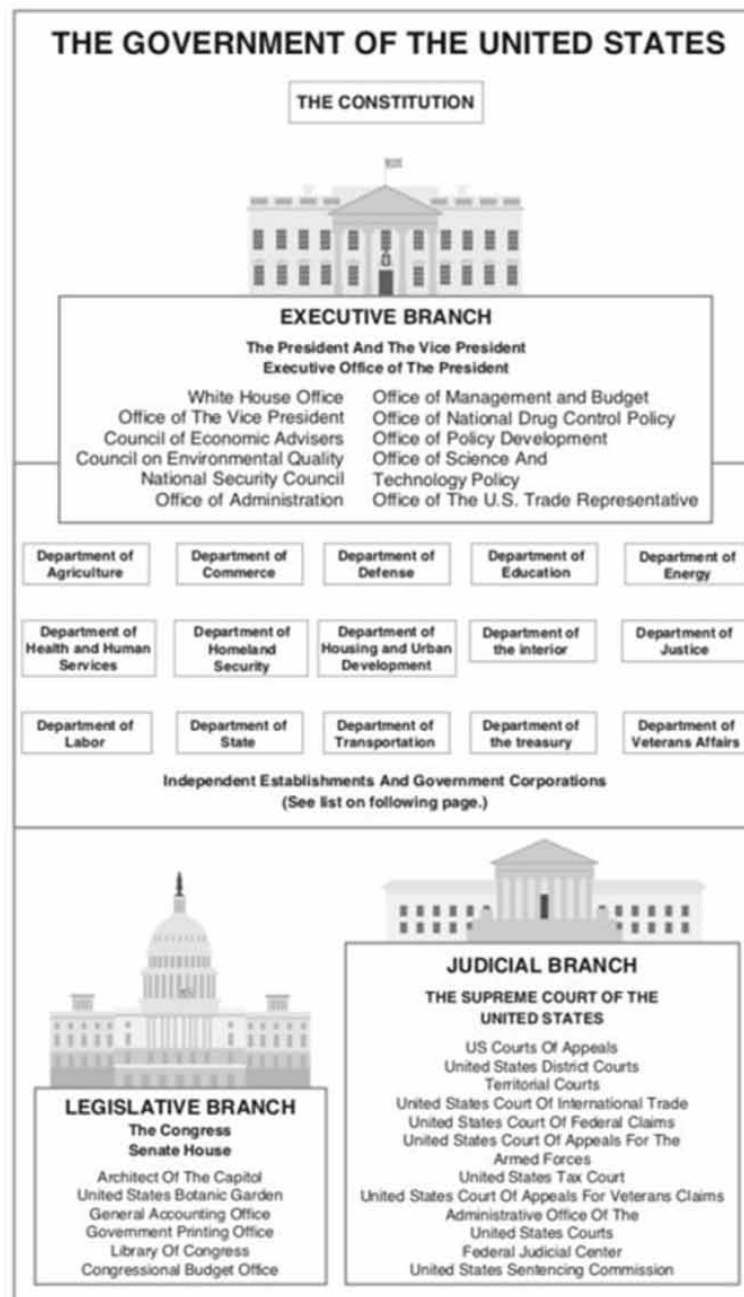
Controversy: Addressing standards for elected office requires Congressional action and collaboration among diverse political interests. In the absent of these standards and collaboration, the economic, social and political aspects are not optimized and are not set on efficient frontier.

Problem: The American Political System is broken and divided and under attack by foreign adversaries. the chapter’s objectives in light of these issues, controversies, and problems

1. Demonstrate how to define what makes a Presidential candidate or incumbent a good one.

The “Background” covered much ground in describing what it takes to be a good president, with a significant omission. That is to consider the entire Federal Government Enterprise and its complexity. No legitimate Commander-in-Chief would approach the job as being a “know-it-all genius.” To lead the nation requires enormous humility and respect for the pluralistic democracy, the rule of law, and its co-equal partners. Divisiveness does not serve the electorate well.

Figure 1. The Government of the United States



When considering a Presidential candidate's resume, ask for evidence of where they obtained knowledge associated with the operation of each Department.

- 1.1 For every Department from which the heads are direct reports to the President, apply the Presidential Tasks.
- 1.2
- 1.3 For every Department identify and define how the organization's systems and operations serve the nation as reflected in essential outcomes.
- 1.4
- 1.5 Define the specific Presidential interactions with the US Congress regarding planning, policy-making, budgeting, implementing laws and regulations, and proposing changes.

The principal knowledge requirements include Constitutional Law and Systems Engineering because laws govern all government enterprise operations and they are implemented in an automated regulatory environment.

Knowledge, Skill, Experience, and Proficiency Requirements are derived from the President's Job Model and from a thorough consideration of all President's past performance.

The operating environment of the US Government is one of collaboration and consensus in a pluralistic structure. The government is a system of systems, governed by laws and regulations where elected officials should possess knowledge of the law and knowledge of systems engineering.

(George & Rodger, 2017) discussed the US presidential candidates and concentrated on relevance of US Constitutional Requirements and relevance of presidential candidate resumes. Education, knowledge, skill, experience, proficiency requirements as well as other qualifications can be summarized as¹

"Education

A president of the U. S. A. should be expected to have graduated, a Doctorate would be highly respected. Should it matter how well the candidate performed in school? We think so.

Curriculum Major Areas of Study

The job of President is Chief Executive of the Federal Government Enterprise that includes all of its departments and agencies. It is not reasonable for a presidential candidate to be an expert in every aspect of the federal government. It is fair to expect candidates to be 1) knowledgeable about the U. S. Constitution and three branches of government, 2) learned about management functions and disciplines, and 3) that candidates are knowledgeable about primary areas of government responsibility such as national defense and homeland security, justice, foreign policy, health, and human services, energy, and the environment for instance.

Foreign policy and public administration are separate specialized fields of study that map to presidential needs.

Skills

Managerial skills in a contemporary paradigm include the following:

- Planning
- Strategizing

- Predicting and forecasting
- Decision making
- Collaborating and facilitating (aka organizing)
- Acquisition, logistics, procurement
- Financing

Affective

“An attribute of the human experience that describes feelings or emotions and sometimes attitudes or values; often used to describe learning objectives or outcomes.

Values include:

- Allegiance to the Constitution and laws and regulations
- Loyalty to party and community
- Appreciation for diversity and collaboration
- Cherishing and relishing the opportunity for bipartisan cooperation

Work History (Sources of applying knowledge, acquiring new knowledge, and applying and developing skills and proficiency)

Military Service: In the instance of military history that is a form of public service

Community: Consideration is given to leadership and issues advocacy, for example Faith, Pays taxes, No legal encumbrance and Marital status.

Presidential Health

Voters need to make a judgment about presidential candidates' health. They can do it blindly or with the candidates' cooperation.

Presidential Character

The importance of presidential character cannot be understated, and there are professional sources for understanding the Presidential Character including the work of (Barber, James David, 1985).

“First, a president’s personality is an important shaper of his presidential behavior on nontrivial matters.

Second, presidential personality is patterned. His character, world view, and style fit together in a dynamic package understandable in psychological terms.

Third, a president’s personality interacts with the power situation he faces and the national “climate of expectations” dominant at the time he serves.

The tuning, the resonance—or lack of it—between these external factors and his personality sets in motion the dynamics of his presidency.

Fourth, the best way to predict a president's character, world view, and style is to see how they were put together in the first place. That happened in his early life, culminating in his first independent political success."

Operating Environment

Commonality

All governments have one thing in common, observes IT guru, Daniel S. Appleton, "They all redistribute wealth." No matter what the form of governance and economy, governments and their leaders make decisions on what and how the wealth of nations will be spent. They make determinations about their discretion to spend based on judgments about the elasticity of their decisions.

In the global economy, trading partners and the aggregate body of global governance determine the accuracy and truthfulness, and the viability of nation-state decisions. The aggregate body of governance can be formal organizations and arrangements of international law and agreements, or it can be informal realized through the behavior of nation states toward one another in trade and defense policies.

Optimizing the Return on National Resources

Work as described in the authors' books, suggests that the ultimate responsibility of governments is to ensure the optimal return on "national" resources. Observe that it not the same as "natural resources," though natural resources are included along with arable land, materials, capital, people, infrastructure and technology assets. What is the optimal ratio of population size and demographics to available and addressable resources? The answer lies deep in the data.

Since humanity on the small planet Earth is dependent on clean air and water, governments must ensure that the resources are protected and provided to citizens. Debate about climate change and the effect of industrialization and overpopulation on the environment all stem from the necessity to balance economic, social, and environmental responsibility in triple-bottom-line accounting.

The Chinese government must weigh the effect of industrial development on the environment just as does the US government. Hazards to local citizens and the cost of healthcare and the resulting calamities must be weighed in both the immediate and long-term.

Opportunities for invention and innovation lies in the pursuit of solutions that stem overpopulation and guide people to live and prosper in the locations and systems that afford a good life for them.

Defining a Good Life

A good life means that people are able to enjoy their needs in conditions of dignity, not living in conditions whereby the only way to satisfy their basic needs is by depriving themselves of their basic freedoms. In a good life, people are able to work in flexible way to sustain themselves and their families, providing adequate food, shelter, clothing and transportation to get to school and work. People require housing which is one of the most important requirements of a decent living and is in fact the basis of family and community stability. People want jobs that provide opportunity to improve their skill, knowledge, and proficiency to the best of their abilities. Impoverishment is unacceptable as that inhibits individuals from optimizing their contribution to society no matter what the form of government and economy.

One important economic indicator that economists consider when analyzing the nation's economic performance is the public debt. When the nation has a considerable amount of debt and when there is a revenue shortfall with mounting deficit, those are significant trouble indications. At the end of the George W. Bush presidential term in 2007, America was facing a financial collapse (James A. George, 2019).

“Financial distress is an extraordinary problem because the nation's security and viability are at stake. Nothing is more important than managing to restore financial integrity because everything else depends upon that.

What are some of the possible causes for a nation's financial stress?

- *Citizen needs exceed the government's capacity to satisfy them.*
- *Government commits to doing more than it has the revenue capacity to achieve.*
- *The nation endures disasters from which resources are insufficient to recover.*
- *The nation embarks on foreign policy that requires military support at a very high cost.”*

In America today (2019), unemployment is low and the GDP is up as is the stock market. Wages are increasing at a slow rate. However, government borrowing and spending results in enormous debt and deficits. The debt is greater than what America produces in a whole year. This high debt-to-gross domestic product ratio tells investors that the country might have problems repaying the loans (Kimberly Amadeo, 2019). How long can government sustain spending before the bubble bursts again?

Private Sector and Corporate Point of View

Corporations in free-market economies aim to optimize profits. They set aside profit to compensate executive leadership and talent, and to reserve capital for continuing improvement, innovation, and invention.

In mixed economies, corporate performance and behavior is regulated because unmitigated capitalism will maximize profits while sacrificing social and Social and natural capital even, the role of government is to create an optimal environment in which they perform while regulating for the good of the state that is their host and for the worker citizens.

The political debate in America is about how much regulation is necessary. It is about the rule of law operative in a global economy.

Looking at corporations as systems and considering their interaction with government systems is essential to corporations optimizing their performance with respect for triple-bottom-line accounting.

A GOOD AMERICAN PRESIDENT

Citizens expect a lot from their Presidents. Understandably, Americans want the President outcome which such as to taking quick action on problems facing the nation and to communicating with the public informing them resolving such problems. Good presidents have outcomes based on a strategic vision and a direction in which they want to lead the country. They are concerned doing tasks that deal with big issues and big challenges, and seek to explain their vision in a way that allows people to understand their circumstances and develop confidence in those proposals. President who succeeds makes the citizens

feel that the nation can succeed (Karl Rove, 2002)². Thus, all Americans participate in measuring the president's performance through feeling the success and achieving that success.

Measuring a Presidential Candidate or Incumbent

In order to know a Presidential candidate or incumbent is a good one; there is a need for measuring the candidate or incumbent. The challenge is to employ advanced information technology to map the ontology and to connect the threads such that information and data are actionable in evaluating and measuring presidential candidates and incumbents.

To arrive at a plausible answer, the authors created a matrix with three major columns:

1. Outcomes
2. Tasks
3. Performance Measures and Metrics

Corresponding to outcomes and tasks is another column called Performance Measures and Metrics. This approach associates "Outcomes" with "Presidential Tasks" and associated "Metrics." Some subtasks are omitted in the chart for brevity.

A President is responsible for on-going government services, systems, and operations as legislated and funded by Congress. Therefore, the first act is to manage the bureaucracy to deliver the required services to the satisfaction of Congress and Citizen Users.

Second, a President may propose changes and improvements, including retiring systems and services. However, that is subject to Congressional approval.

The question is, how does the US government measure the performance of Departments and Agencies that are responsible for services delivery? Since the President leads the Executive branch in staffing all of the departments and agencies, one of the essential tasks is staffing.

Voters Vet Candidates

As non-profit corporation political parties should be accountable and liable for vetting candidates and incumbents bearing their brand. However, the judiciary may not agree.

"Update: A federal judge on August 28. The court recognized that the DNC treated voters unfairly, but ruled that the DNC is a private corporation; therefore, voters cannot protect their rights by turning to the courts:

"To the extent Plaintiffs wish to air their general grievances with the DNC or its candidate selection process, their redress is through the ballot box, the DNC's internal workings, or their right of free speech — not through the judiciary." (Independent Voter, 2017)

Let the voters beware.

Citizens may not have the time, skill, or knowledge to properly vet candidate resumes. They are dependent on political parties, and other third-parties for assistance. Needed are technology inventions such as apps to help them manage the data and to answer question.

(The League of Women Voters, 2019) is one example of a third-party source of candidate information.

Table 2. President's Task Performance, Measures and Metrics

Outcomes	President's Task Performance	Measures & Metrics
1. Human Rights Assured	Task 2: Develop the President's management agenda and budget and reconcile with Congress Evidence of specific agenda and compliance with international agreements and commitments.	100% Universal Declaration of Human Rights
2. Constitutional/ Bill of Rights Allegiance	Task 2: Develop the President's management agenda and budget and reconcile with Congress Subtask 2.1: Assume responsibility for the legacy agenda and make adjustments to align with the President's management agenda or equivalent Subtask 2.2: Work with cabinet heads to develop performance plans and schedules Subtask 2.3: Work with Congressional leadership and collaborate to implement the nation's strategy, plans, and programs Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation Subtask 3.1: Propose bills to Congress Subtask 3.2: Consult and advise Congress Subtask 3.3: Collaborate with industry and business leaders in the development of policies and regulations of all kinds Subtask 3.4: Approve or veto legislation Subtask 3.5: Request a declaration for war from Congress Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations Subtask 4.1: Review and evaluate programs that include new acquisitions and on-going operations Subtask 4.2: Evaluate programs including legacy processes and engineer new ones for accomplishing and producing required and promised outcomes Subtask 4.3: Continuously improve Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	100% The US Department of Justice Strategic Plan is out of date and noncompliant. The US President has allegedly obstructed justice and awaits Congressional action. The US President dishonors subpoenas from Congress by directing that they be ignored by the Executive branch.
3. Rule of Law Followed	President's actions such as Executive Orders comply with the law.	100% Measures: Court Findings for and against the President. The US Department of Justice Strategic Plan is out of date and noncompliant. Many of President Trump's EOs have been stricken by the court as being illegal and unenforceable.
4. Economic Security	Task 3: Initiate and approve legislation Subtask 3.1: Propose bills to Congress Subtask 3.2: Consult and advise Congress Subtask 3.3: Collaborate with industry and business leaders in the development of policies and regulations of all kinds Subtask 3.4: Approve or veto legislation Subtask 3.5: Request a declaration for war from Congress Task 4: Implement plans and manage on-going operations Subtask 4.1: Review and evaluate programs that include new acquisitions and on-going operations Subtask 4.2: Evaluate programs including legacy processes and engineer new ones for accomplishing and producing required and promised outcomes Subtask 4.3: Continuously improve	100% Measures and Metrics Correspond with all appropriate Departments and Agencies plans, programs, and budgets. According to Budget Request Annual Performance Plan and Reports of "The Government Performance and Results Act Modernization Act (GPRAMA) of 2010 requires Treasury and other federal agencies to formally monitor and review organizational performance through strategic reviews, the annual performance plan, and the annual performance report." "The Congressional Justification of Appropriations (CJ) reflects the President's Budget request for the Department of the Treasury and is prepared in accordance with OMB Circular A-11. The CJ includes agency priorities, requested budget levels, and performance plans and reports (in accordance with GPRAMA)."

continued on following page

Table 2. Continued

Outcomes	President's Task Performance	Measures & Metrics
5. Sustainable Economy	Task 2: Develop the President's management agenda and budget and reconcile with Congress Subtask 2.1: Assume responsibility for the legacy agenda and make adjustments to align with the President's management agenda or equivalent Subtask 2.2: Work with cabinet heads to develop performance plans and schedules Subtask 2.3: Work with Congressional leadership and collaborate to implement the nation's strategy, plans, and programs Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	100% Measures and Metrics Correspond with all appropriate Departments and Agencies plans, programs, and budgets. There is no legal requirement for a sustainable economy.
6. Environmental Security	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	100% Clean Air and Clean Water "Planning, Budget, and Results EPA's planning, budget, and results activities enable EPA to carry out its mission through: · coordinating EPA planning efforts and preparing EPA's Strategic Plan; · developing and managing EPA's budget; and · reporting on EPA's financial and performance results." (U.S. Environmental Protection Agency, 2019) "The Clean Water Act (CWA) establishes the basic structure for regulating discharges of pollutants into the waters of the United States and regulating quality standards for surface waters. ... Under the CWA, EPA has implemented pollution control programs such as setting wastewater standards for industry. Mar 11, 2019" (U.S. Environmental Protection Agency, 2019) "The Clean Air Act (42 U.S.C. § 7401) is a United States federal law designed to control air pollution on a national level. ... The first federal legislation to actually pertain to "controlling" air pollution was the Clean Air Act of 1963."

continued on following page

Table 2. Continued

Outcomes	President's Task Performance	Measures & Metrics
7. National Military Security	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	100% Measures and Metrics Correspond with all appropriate Departments and Agencies plans, programs, and budgets. Observe that the GAO evaluates GPRA implementation by department and agency and issues a report. "Pursuant to a legislative requirement, GAO reviewed the Department of Defense's (DOD) implementation of the Government Performance and Results Act of 1993 (GPRA). GAO noted that: (1) the implementation of GPRA within DOD is in its initial stages and many implementation strategies and key approaches need to be developed, but DOD has made progress at various organizational levels in implementing GPRA or strategic planning and performance measurement initiatives consistent with GPRA; (2) DOD has developed strategic goals and objectives and is refining a set of performance measures to meet the planning requirements of GPRA; (3) some subordinate DOD organizations, building on their experience as GPRA pilots, are developing or improving strategic planning and performance measurement systems; some other defense organizations are developing or improving such systems as well; (4) GAO's work suggests and DOD officials agree that DOD's effectiveness in implementing the requirements of GPRA may be improved by better hierarchical linking of goals and performance measures; (5) for example, GAO recently examined the degree to which DOD's logistics strategic plan provides an integrated logistics roadmap for the department to support its warfighting strategy; (6) while the services' strategic plans and initiatives generally support implementation of the DOD plan, their goals, objectives, and strategies are not always directly linked; (7) DOD indicated that it will ensure that the next edition of the DOD Logistics Strategic Plan includes specific guidance to require the services to link their goals to DOD's; (8) this is important because, without this goal alignment, DOD may have difficulty meeting its department-wide logistics goals, which are to reduce logistics response time, develop a seamless logistics system, and streamline the logistics infrastructure; (9) GAO's recent work also found that DOD's strategic information resources planning effort does not appear to link its information resources management systems development with recent initiatives focusing on consolidating or privatizing various areas of logistics operations; (10) if not linked, DOD could end up spending millions of dollars on systems designed to support functions that it might not plan to do or in organizations that might be eliminated; (11) DOD stated that it is addressing these concerns by preparing a new logistics business systems strategy; (12) GAO's review of DOD's ongoing GPRA implementation efforts suggests that hierarchically linked goals and measures will be crucial for success; (13) DOD must ensure the services implement GPRA at their level and provide goals that link DOD's department strategic goals with those of lower level components; and (14) some of the officials assigned GPRA liaison responsibilities told GAO they did not know what, if anything, they should be doing to promote GPRA principles within their organizations." (U.S. Government Accountability Office, 1997)
8. Education	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	

continued on following page

Table 2. Continued

Outcomes	President's Task Performance	Measures & Metrics
9. Homeland Secured	Task 2: Develop the President's management agenda and budget and reconcile with Congress Subtask 2.1: Assume responsibility for the legacy agenda and make adjustments to align with the President's management agenda or equivalent Subtask 2.2: Work with cabinet heads to develop performance plans and schedules Subtask 2.3: Work with Congressional leadership and collaborate to implement the nation's strategy, plans, and programs Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	100%
10. Full Employment	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	100%
11. Upward Workforce Mobility	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	
12. All Citizens minimally sustained	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	No Poverty

continued on following page

Table 2. Continued

Outcomes	President's Task Performance	Measures & Metrics
13. Social Security for all Disabled and Retired	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	
14. Free Nation	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	
15. Low Crime	Task 2: Develop the President's management agenda and budget and reconcile with Congress (Subtasks omitted) Evidence of specific agenda and compliance with international agreements and commitments. Task 3: Initiate and approve legislation (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes. Task 4: Implement plans and manage on-going operations (Subtasks omitted) Evidence of specific plans, programs, initiatives and orders resulting in specific measurable outcomes.	
16. Government Services Delivered	Task 1: Planning, staffing, organizing, and scheduling Presidential work and government functions Subtask 1.1: Recruit and staff the cabinet and department and agency appointments Subtask 1.2: Conduct cabinet meetings to develop and implement strategies and policies for accomplishing the nation's workload and issues Subtask 1.3: Define the nation's outcomes and priorities for each major department and agency to produce the nation's strategic plan	100% As legislated and required by law. Performance is measured by each department and agency in accordance with Congressional Oversight and Executive Branch scrutiny.

“We are all equal at the ballot box, but only if we vote. We engage millions of voters every year ensuring Americans have the information they need to participate in elections that determine our future.” Elections.

Evaluating Candidates and Incumbents

Managing the large and complex US Federal Government Enterprise requires collaboration and consensus among three co-equal branches of government. The President must possess leadership and behavioural traits for achieving results in a bipartisan environment.

Being a CEO of a private corporation is considerably different from being a US President because the aim is not profit-driven, and the relationship between the chief executive and the other branches of government is dissimilar to that of the relationship with a Board of Directors.

In some sense, Citizens are considered shareholders in the US Government because they pay for it with their taxes, and they elect leadership and representatives. Citizens receive a return on their investment as realized in services and outcomes.

Citizens expect and demand loyalty from the President to the rule of law, the US Constitution, and executing the Oath of Office with high integrity. Lying by a federal government official is against the law and therefore a President must guard their words as being honest and truthful. Contradictory is the fact that a sitting President cannot be prosecuted for crimes, and can only be impeached by Congress.

Solutions and Recommendations

The order of complexity inherent in a democratic republic and its associated Federal Government Enterprise is enormous. However, with the aid of technology and improved systems and tools, citizen voters can improve their choices and decision making.

The greatest challenge is to inspire elected officials to redouble their focus and attention to the priorities and essentials of good governance. The President is the leader in ensuring a good life for all citizens.

The specific issues and needs change in response to the global environment. America is a member of the global community from which isolation is not an option. A President must apply collaboration skills broadly in conducting foreign policy and to secure the nation and to contribute to global peace, prosperity, and security.

Political science, law, economics, and systems engineering intersect as disciplines for continuous improvement and that is what the authors wish to bring to light.

The overarching legislation governing performance management in the Federal government is The GPRA Modernization Act of 2010.

“The GPRA Modernization Act modernizes the Federal Government’s performance management framework, retaining and amplifying some aspects of the Government Performance and Results Act of 1993 (GPRA 1993) while also addressing some of its weaknesses. GPRA 1993 established strategic planning, performance planning and performance reporting as a framework for agencies to communicate progress in achieving their missions. The GPRA Modernization Act establishes some important changes to existing requirements.

The purposes of the GPRA Modernization Act are to:

- *Improve the confidence of the American people in the capability of the Federal Government, by systematically holding Federal agencies accountable for achieving program results;*
- *Improve program performance by requiring agencies to set goals, measure performance against those goals and report publicly on progress;*
- *Improve Federal program effectiveness and public accountability by promoting a focus on results, service quality and customer satisfaction;*
- *Help Federal managers improve service delivery, by requiring that they plan for meeting program goals and by providing them with information about program results and service quality;*
- *Improve congressional decision-making by providing more information on achieving statutory objectives and on the relative effectiveness and efficiency of Federal programs and spending;*
- *Improve internal management of the Federal Government; and*
- *Improve usefulness of performance and program information by modernizing public reporting”* (Bureau of Reclamation, 2010)

The question today (2019), is the Executive Branch under incumbent Donald Trump, following the law? Is Congress enforcing the law with oversight?

FUTURE RESEARCH DIRECTIONS

Because democracy and its institutions are intended to be accessible and addressable to average citizens, it is essential to develop education curricula and tools, including software applications that help voters increase their engagement and interaction in government processes. Herein lies much opportunity to break new ground with inventions and innovations.

The authors suggest establishing standards for political party platforms and for Presidential candidates for mapping their manifestos to the platforms. Also recommended is creating an “app” to assist citizen voters in evaluating candidate resumes and credentials.

CONCLUSION

This chapter raises more questions though has begun to answer many to initiate further research and actions by citizens, political parties, and the media to help ensure that the democratic republic will endure against threats to equality and freedom. Perhaps the most significant risk is from complacency and cynicism. Democracy is a living institution and is adaptive to change and improvement. It depends on intellectual investment for its survival and capacity to thrive.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

- Amadeo, K. (2019). The US Debt and How It Got So Big: Five Reasons Why America Is in Debt. <https://www.thebalance.com/the-u-s-debt-and-how-it-got-so-big-3305778>
- Barber, J. D. (1985). *The presidential character: predicting performance in the White House* (3rd ed.). Prentice-Hall.
- Bolívar. (2015). Risk Factors and Drivers of Financial Sustainability in Local Government: An Empirical Study. *Journal of Local Government Studies*, 42(1), 29–51. doi:10.1080/03003930.2015.1061506
- Brichacek, A. (2019, September 8). Six ways the media influence elections. Retrieved from University of Oregon School of Journalism and Communication: <https://journalism.uoregon.edu/news/six-ways-media-influences-elections>
- Bureau of Reclamation. (2010, December 31). Retrieved from U.S. Department of Interior: Reclamation Managing Water in the West: <https://www.usbr.gov/gpra/#Daily>
- Callen, T. (2018, December 18). Gross Domestic Product: An Economy's All: Finance & Development. Retrieved from The International Monetary Fund (IMF): <https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm>
- Cork. (2007). Ecosystem Services and Australian Natural Resource Management (NRM) Futures. <https://www.environment.gov.au/system/files/resources/11543d24-9f2c-44ee-a52c-83dfa6adb7d9/files/ecosystem-services.pdf>
- George, J. A. (2019). A Good Life for All Citizens for an Entire Lifetime. In *A President's Manifesto. Indiana of Pennsylvania*. James A.
- George & Rodger. (2017a). How to Select an American President. Bloomington: Archway Publishing. <https://www.archwaypublishing.com>
- George & Rodger. (2017b). Regenerating America with Sustainable Economics. Bloomington, IN: Archway Publishing. <https://www.archwaypublishing.com/Bookstore/BookDetail.aspx?Book=763073>
- Goldsmith, P. (2019, September 8). Pluralism: Levels: AS, A Level Exam boards: AQA, Edexcel, OCR, IB. Retrieved from tutor2u: Politics: <https://www.tutor2u.net/politics/reference/pluralism>
- Gretchen, C. (1997). *Nature's Services: Societal Dependence on Natural Ecosystems*. Island Press., <http://willsull.net/la370/resources/Module-2/Daily.pdf>
- Hilton. (2007). Human rights policy statement. U.S. Virgin Islands. <https://ir.hilton.com/~media/Files/H/Hilton-Worldwide-IR-V3/committee-composition/he-human-rights-policy-2017.pdf>
- History, U. S. org. (2019, September 8). American Political Culture: American Government Online Textbook. Retrieved from U.S. History. Org: Independence Hall Association: <https://www.ushistory.org/gov/4a.asp>
- Ignatov, D. (2017, March 8). Computer Science > Information Retrieval: Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields. Cornell University. arXiv. orgLogin: [arXiv.org > cs > arXiv:1703.02819](https://arxiv.org/abs/1703.02819)

Matsa & Shearer. (2018, September 10). News Use Across Social Media Platforms 2018: Most Americans continue to get news on social media, even though many have concerns about its accuracy. Retrieved from Pew Research Center: Journalism & Media: <https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/>

National Priorities. (2016, December 31). Fighting for a U.S. federal budget that prioritizes peace, economic security and shared prosperity: Federal Revenue: Where Does the Money Come From. Retrieved from National Priorities: <https://www.nationalpriorities.org/budget-basics/federal-budget-101/revenues/>

Office of Personnel Management (OPM). (2019, September 9). Policy: Senior Executive Service: lead America's workforce. Retrieved from The U.S. Office of Personnel Management (OPM): <https://www.opm.gov/policy-data-oversight/senior-executive-service/>

Pettinger, T. (2011, February 22). Genuine Progress Indicator GPI v GDP. Retrieved from Economics help blog: <https://www.economicshelp.org/blog/2666/economics/genuine-progress-indicator-gpi-v-gdp/>

Pooja. (2019, September 8). What are the main Objectives of Political Engineering? Retrieved from Political Science Notes: <https://www.politicalsciencenotes.com/articles/what-are-the-main-objectives-of-political-engineering/418>

Rove, K. (2002, November 13). What Makes a Great President. Retrieved from Columbia College of Arts and Sciences: History News Network: <https://historynewsnetwork.org/article/1529>

Techopedia. (2019, September 10). Concept of Operations (CONOP). Retrieved from Techopedia Definition: <https://www.techopedia.com/definition/24826/concept-of-operation-conop>

The League of Women Voters. (2019, September 7). Elections. Retrieved from The League of Women Voters: <https://www.lwv.org/elections>

United Nations. (1997). *Glossary of Environment Statistics: Series F, No. 67*. United Nations.

U.S. Bureau of Economic Analysis. (2019, August 29). Gross Domestic Product. Retrieved from Bureau of Economic Analysis: <https://www.bea.gov/data/gdp/gross-domestic-product>

U.S. Department of the Treasury. (2019, September 7). Budget Request/Annual Performance Plan and Reports. Retrieved from U.S. Department of the Treasury: <https://home.treasury.gov/about/budget-financial-reporting-planning-and-performance/budget-requestannual-performance-plan-and>

U.S. Environmental Protection Agency. (2019, September 7). Laws & Regulations: Summary of the Clean Air Act. Retrieved from United States Environmental Protection Agency "EPA": <https://www.epa.gov/laws-regulations/summary-clean-air-act>

U.S. Environmental Protection Agency. (2019, September 7). Laws & Regulations: Summary of the Clean Water Act. Retrieved from United States Environmental Protection Agency: EPA: <https://www.epa.gov/laws-regulations/summary-clean-water-act>

U.S. Environmental Protection Agency. (2019, September 7). Planning, Budget, and Results. Retrieved from United States Environmental Protection Agency: <https://www.epa.gov/planandbudget>

U.S. Government Accountability Office. (1997, January 31). DOD's GPRA Implementation: GAO reviewed the Department of Defense's (DOD) implementation of the Government Performance and Results Act of 1993 (GPRA). Retrieved from U.S. Government Accountability Office (GAO): <https://www.gao.gov/products/NSIAD/GGD-97-65R>

Voter, I. (2017, May 2). DNC to Court: We Are a Private Corporation With No Obligation to Follow Our Rules. Retrieved from Independent Voter News: <https://ivn.us/posts/tremors-in-the-campaigns-point-to-breakup-of-the-major-parties>

Walzer, M. (2007, November 31). Pluralism and Democracy. Retrieved from The Atlantic Magazine: <https://www.theatlantic.com/magazine/archive/2007/11/pluralism-and-democracy/306321/>

KEY TERMS AND DEFINITIONS

Collaboration: Working together to create something that is better than going it alone. In a democratic republic, representatives from different political parties collaborate to attend to the needs of the nation, where getting things done in a bipartisan manner is more important than political ideology.

Concept of Operations (CONOPS): The concept of operations (CONOPS) of the American Political System begins with the US Constitution. Though there is also the notion of political parties and their role in offering ideological brands based on or relating to a system of ideas and ideals, especially concerning economic or political theory and policy. The extent to which party brands align with what the "Framers" had in mind is open for debate.

Ecosystem Services: Conditions and processes through which natural ecosystems and the species that make them up sustain and fulfil the human life. Thus, provide such things like agricultural produce, timber, and aquatic organisms such as fishes and crabs (Daily, 1997).

Group Knowledge Model: In this chapter, group knowledge model refers to the process of information about the subject to socialize the discussion and to enhance individual understanding.

Job Model: A job model is a technique for defining the work performed from a specific position in an organization. The author learned the approach while being employed by AT&T and found it compatible with process modeling techniques such as the Integrated Definition Language (IDEF) employed extensively by the US Department of Defense and other departments and agencies. Work is defined as tasks and subtasks that are akin to activities and arrayed in a logical hierarchy.

Operating Environment: In computer technology terms, an operating environment or integrated applications environment is the environment in which users run application software. In this chapter, the operating environment encompasses all of the processes and systems as legislated and funded by Congress and managed by the Executive branch under oversight from Congress. The technical infrastructure is computer and communications technology.

Optimize: To optimize in this chapter means making the highest yield and best use of government resources in the process of attending citizen needs. Optimizing requires trade-offs and deliberate consideration for risks and benefits toward producing required outcomes and associated metrics.

Optimized Sustainment Model (OSM): The authors proposed the Optimized Sustainment Model in the book *Regenerating America with Sustainable Economics* by James A. George and James A. Rodger © 2017 Archway Publishing. The idea is to develop business rules and regulations that support the application of triple-bottom-line accounting methods to measuring corporate performance. The idea is to balance rewards for economic, social, and environmental responsibility. It is a work in progress for which academics are encouraged to contribute. The model is suggested as an alternative to capitalism.

Private Sector: The part of the national economy that is not under direct government control, though for which corporations and citizens are regulated.

Public Sector: The part of an economy that is controlled by the government and funded by taxpaying citizens who control the government.

ENDNOTES

¹ These skills and others are detailed in the book.

² Mr. Rove was President Bush's chief political advisor.

APPENDIX

Table 3. Presidential Candidate Evaluation Matrix

Evaluation Criterion	Values	Score
Qualifications Required by Law: As directed by the Constitution, a presidential candidate must be a natural born citizen of the United States, a resident for 14 years, and 35 years of age or older.	1	1
Home State/Residence: Some might favor a larger state over a smaller one based on affinity with governing the nation at large. Values range 1-3.	1-3	
College Graduate BS/BA	1	
GPA: H-2, M-1, L-0	0-2	
Academic Honors and Distinctions	1	
Post Graduate – Masters	2	
PhD	3	
Occupation/Vocation		
Lawyer	3	
CPA	2	
Professional i.e. MD	2	
Physical Health	1-3	
Mental Health	0-1	
Intelligence Quotient	1-3	
Affected Class (Race, Gender)	1-3	
Character	0-2	
Leadership	1-3	
Continuing Education/Training	1	
Knowledge about Government	1-3	
Knowledge about the Job	1-3	
Relevant Skills	1-3	
Life History	0-2	
Manifesto/Platform/Agenda	1-3	
Relationships	1-3	
Political Party Affiliation	0-2	
Values	1-3	
Military Experience	1	
Veteran	2	
Officer	2	
NCO	1	
None	0	
Public Office: Vice President	5	
Public Office: Judge	1	
Public Office: Dept. Secretary	3	
Governor	4	
US Senator	3	
US Member of the House of Representatives	3	

continued on following page

Table 3. Continued

Evaluation Criterion	Values	Score
Mayor: (Size of municipality matters)	1-3	
State Legislator	1	
Other Public Office	1	
Private Sector CEO/President (Size and nature of business/industry matters)	1-3	
Director/Manager	1	
Inventor (Patents)	1-2	
Entrepreneur	1	
Memberships	1	
Religion: Matters, Irrelevant, Private (Stated position)	1	
Incumbency	0-10	
Possible Score Range (<61-94)	61-94	

The Role of DPPs in Promoting Local Government–Citizen Collaboration and Participation: The Case of “Baladiaty”

Khaled Tamzini

 <https://orcid.org/0000-0001-8397-2769>

LaREMFiQ Laboratory, IHEC of Sousse, University of Sousse, Tunisia

Ynes Hafi

ARSELA, Tunisia

Achref Ben Ouannes

ARSELA, Tunisia

Roula Borhani

OMEDRH, Tunis, Tunisia

INTRODUCTION

Tunisia is considered as an emerging economy, a very young and nascent democracy not only in Africa but also in all the Arabic Word. In fact, since the 2011 revolution, Tunisia has undergone a rapid political transformation to democracy. After six decades of autocracy, it has witnessed four consecutive free and fair elections, including the 2011 constituent assembly elections, the 2014 parliamentary and presidential elections, and the 2018 municipal elections. It approved a new constitution in 2014 that is progressive even by Western standards, enshrining not only freedom of religion but also conscience (permitting atheism) and mandating not just gender equality but an active commitment by the state to ensuring it. Tunisia has also passed progressive laws countering violence against women and racial discrimination (Grewal, 2019).

But eight years into democracy, Tunisians have become frustrated with its failure to deliver economically. The unemployment rate, which had remained at a steady 12 to 13 percent through the late 2000s, jumped to 18 percent after the 2011 revolution and remains at 15 percent today. The rate of inflation, 3 to 4 percent prior to the revolution, has doubled to 7.4 percent. Receipts from international tourism, the third-largest sector of the Tunisian economy, totaled \$3.9 billion in 2009 and have dropped to just \$1.7 billion today. By nearly all metrics, the economic situation is even worse than that which prompted the 2011 revolution (Grewal, 2019).

Accordingly, Tunisians have become increasingly disillusioned with democracy as a solution to their day-to-day problems. Nationally representative survey data from the Afrobarometer help to capture this disillusionment over the course of Tunisia’s transition. While 70 percent of Tunisians in 2013 agreed that “democracy was preferable to other forms of government,” only 46 percent agreed last year (Grewal,

DOI: 10.4018/978-1-7998-3479-3.ch108

2019). Moreover, Tunisians no longer trust their parliaments, municipal governments, and political parties. They have lost the commitment and engagement of their citizens.

In this context, what could be done to deal with this dramatic situation?

The literature review in the public management field suggests the use of digital tools as one of the multiple answers to this question. In other words, how digital tools can be used by parliaments, municipal governments, and political parties to engage citizens in the reconquest of democracy?

Desouza & Bhagwatwar (2012); Ertiö (2015); Williamson & Parolin (2013); and Zavattaro & Sementelli (2014) have claimed that we need more digital tools for actual collaboration and citizen engagement in local government activities (Falco & Kleinhans, 2018).

For Simon et al. (2017), “*the practice of democracy using digital tools and technologies*” is called “*digital democracy*”. Therefore, among these digital tools, the authors distinguish the Digital Participatory Platform (DPP) as a specific kind of collaborative social media. For Falco and Kleinhans (2018) DPPs include all the features proper to ‘conventional’ social media (such as Facebook and Twitter): they are based on Web 2.0, allowing for user-generated content, and sharing of such content. However, compared to conventional social media, DPPs also include different and more elaborate technological features.

The implementation of a Digital Participatory Platform (DPP) takes into account the mandatory use of issues and challenges that local governments face in fostering online and offline citizen engagement (Falco and Kleinhans, 2018). In fact, what any local government, especially municipality aspire or look for is not only implementing DPP, but also how this later should be designed to foster citizen engagement and democracy.

Thus, the aim of this paper is to present the implementation of a pilot digital participatory platform (DPP) called “Baladiaty” in order to foster local government-citizens collaboration and participation in Tunisia. “Baladiaty” means in the Arabic language “My municipality”. “Baladiaty” is a DPP designed, developed and commercialized by ARSELA in 2019, a Tunisian start-up located in the governorate of Sousse.

This paper will contain two main parts. The first is considered a theoretical part. The authors will focus on the two main terms and concepts of this research, namely: digital democracy, and DPP. This first part will focus on the definition of the digital participatory platform (DPP), and its role in promoting participatory democracy, citizen engagement and their ubiquitous engagement throughout the open government concept. The second part would be considered as a case study presenting a recently designed pilot DPP “Baladiaty” in Tunisia. This paper could be considered theoretical by focusing on the presentation of a pilot DPP in Tunisia, a nascent democracy.

BACKGROUND: TERMS AND CONCEPTS

In the following, the authors will focus their literature review on the main concepts of this chapter: democracy, e-democracy, and digital participatory platform (DPP). This literature review, based on recent publications, will cover the specificities of all these fundamental concepts: definitions, objectives, dimensions, forms, and models, etc.

I. Democracy

Before going further on the presentation of e-democracy and DPPs, the authors chose to focus on certain definitions of democracy, its forms, and models in order to clarify their foundation.

1. Definitions

For several researchers and patricians (Simon et al., 2017; Abu-Shanab, 2015; Dalton et al., 2007) democracy is not easy to define. Thus, before moving on the presentation of some contemporary definitions of democracy, the authors will first focus on these two antagonistic currents of thought in defining democracy: the elitists, and the participationists. In fact, among the proponents of the first current of thought called “elitists”, Lundström (2004) distinguished Schumpeter (1952), Sartori (1974), Downs (1957), and Dahl (1971) who “*recommend an elite-dominated form of democracy, on the grounds that such a system serves to promote social peace, stability, and efficiency. They regard a democratic competition between elites as more realistic and appropriate than a broad democratic participation in the exercise in political power*” (Lundström, 2004, p. 5). For instance, Schumpeter (1952, in O'Donnell, 1999) defined democracy as “... *a political method ... a certain type of institutional arrangement for arriving at political - legislative and administrative – decisions. (...) that institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote*” (p. 11),

Unlike the elitists, the participationists (Mill, 1983; Habermass, 1988; Pateman, 1970) “*recommend participatory democracy, on the grounds that this form of government fosters a sense for the common good, personal development, and equality. Political participation, they aver, brings good consequences*” (Lundström, 2004, p. 5). Thus, Pateman (1970 in Floridia, 2013), to define democracy, put forward the concept of “participation”. She stated, “*In the participatory theory, ‘participation’ refers to (equal) participation in the making of decisions, and political equality refers to equality of power of determining the outcome of decisions* (Pateman, 1970, p. 43)”. In fact, as stated by Floridia (2013) to explain the view of Pateman (1970) “*Equality of the power that individuals exercise in decision-making is the key of this understanding of democracy and participation*” (p. 47).

More recently, Simon et al. (2017) stated, “*democracy is a cluster of practices, structures, institutions, and movements. It's an assembly of many different elements – and it's the combination or totality of those elements that we understand to mean democracy*” (p. 11).

Abu-Shanab (2015), links “democracy” and its prosperity to the political well of the country. In fact, the greater the trust between government and citizens (by allowing citizens to participate in public policy processes and promoting open and accountable government), the greater the country's political well-being (Kneuer, 2016).

For Dalton et al. (2007), defining “democracy” implies taking into account three major alternatives: institutions and procedures, freedom and liberties, and social benefits. These constitute “the primary substantive choices in defining democracy” (p. 145).

As argued by Dalton et al. (2007), institutions and processes of representative government are considered as one of the pillars to define “democracy” and for Dahl et al. (1971 in Dalton et al., 2007) they are equated with “democracy”. Indeed, for the latter “*if citizens can participate equally in free and fair elections, and if elections direct the actions of the government, then the standards of democracy are met. Indeed, Freedom House rankings and other democracy indicators often treat free and fair elections as a defining element of democracy*” (p. 143).

The second pillar to define “democracy” is called “freedom and liberties”. For Dalton et al. (2007) they constitute the outcomes of “democracy”. In fact, its goal is to reach freedom and liberty for all citizens through democratic institutions. Moreover, when one analyzes the four core democratic values (political liberties, participation rights of citizens, equal justice before the law, and equal rights for women) listed by Diamond (1999, in Dalton et al., 2007), the notions such as rights, liberties, and freedom are ubiquitous.

The results of a survey undertaken by Bratton et al. (2005) in Africa and Latin America have found that references to liberty, freedom, and equality were the most common answers in defining the meaning of democracy. Likewise, for the results of another survey undertaken in five East European nations, Simon (1998, in Dalton et al., 2007) have found that liberty and basic rights were the first answer given by a majority of the public.

Social benefits are considered as the third pillar of democracy's definition. In nascent democracies with a social-democratic conception, social services ensure well-being and social peace (Bratton et al., 2005). In developing countries, economic benefits and a higher standard of living take precedence over social benefits. The results of McIntosh and Abele (1993 in Dalton et al., 2007) survey undertaken in East European countries, show that most people equated democracy with the economic values of prosperity, equality, and security that they considered most important to their country's democratic development. *"In contrast, the same study found that citizens in Britain, France, and West Germany emphasized the political values of political freedom, party competition, and a fair justice system"* (Dalton et al., 2007, pp.144-145).

2. Forms and Models of Democracy: Representative, Participatory (Deliberative or Collaborative), and Direct Democracy

Defining « democracy » cannot be done without introducing its forms and models. In this paper, the authors will use the terms forms and models of democracy interchangeably.

The literature review distinguishes three main models and forms of democracy: representative, participatory (deliberative or collaborative), and direct democracy. The main difference between these forms of democracy is located in the nature of the decision-making process (or the democratic process). In fact, the old conception of the democratic process is based on the inability to give the chance to all public members to interfere in the political decision-making process (Clive, 2012). This is called: "representative democracy".

Representative democracy, or representative government is a political system in which a restricted assembly is recognized as having the right to represent a people, a nation or a community. The will of the citizens is expressed through representatives who establish the laws (legislative power) and enforce them (executive power).

At present, all democratic states are representative democracies, in which laws are elaborated by representatives elected by the people or by the government, and voted by parliament. However, representative democracy can be complemented by mechanisms of "direct democracy", or basic democracy (Pohoryles, 2017), so-called because they directly give decision-making power to citizens in certain circumstances.

Direct democracy makes its role of a sovereign to the population by allowing it to take certain decisions itself, in particular by the technique of "the referendum" or that of "the popular consultation". It is based on the principle that the population as a whole will make the best decision via a popular vote, is directly concerned by the consequences of its choice: it is supposed to know, better than the elected, where its will and its interests reside (Pohoryles, 2017).

Participatory democracy can be reduced to the set of mechanisms and procedures that increase the involvement of citizens in decision-making and are based on the shortcomings of representative democracy and its limits in the involvement of citizens in decision-making, implementation, and control. For Pohoryles (2017), participatory democracy limits the extent of public participation and quite often leads to decisions that are not legally binding for the government or parliament.

Simon et al. (2017), distinguished between deliberative democracy and collaborative democracy as two forms of participatory democracy. The main idea of deliberative democracy, inspired by the theories of John Rawls and Jürgen Habermas, is that a political decision is truly legitimate when it proceeds from the public deliberation of equal citizens. In relation to participatory democracy, it emphasizes the need for well-argued debates among citizens. The collaborative democracy is defined as a political system where each citizen can propose his ideas without an intermediary, that is to say without delegation of power to specialists, experts or political representatives, to participate directly in the elaboration of a common political project and involve in decisions.

These two forms of participatory democracy could be considered as the foundations of modern democracy called e-democracy or digital democracy.

II. E-democracy

In this section, the authors will introduce “e-democracy” by defining it, enumerate its typology, and conclude with its dimensions.

1. Definitions

The literature review related to political sciences is very rich in terms of definitions of e-democracy, also called digital democracy. However, some researchers agreed that there is no common definition of e-democracy (Kneuer, 2016; Simon et al., 2017).

Simon et al. (2017) defined e-democracy as *“the practice of democracy using digital tools and technologies”*. But at the same time, they distinguished between “minimalist” and “maximalist” definitions of digital democracy. In fact, *“the former focuses on giving citizens access to governmental information and enabling them to interact with government through, for example, online consultations and transactional services online. The latter envisages a more participatory role for citizens, enabling them to collaborate with government officials as well as make their own decisions about how they and their local communities are governed”*. (Tuzzi et al., 2007 in Simon et al., 2017, p. 11).

For Kneuer (2016) digital democracy is *“an overarching concept, namely the use of ICT by political actors (government, elected officials, media, political/societal organizations, and citizens) within political and governance processes in today’s representative democracy”* (p. 669). For some other authors, e-democracy is equated with citizen participation (Manoharan and Holzer, 2012, in Kneuer, 2016).

Kies et al. (2004, in Abu-Shanab, 2015) defined “e-democracy” as *“the use of ICT tools to empower citizens and enable them to put politicians accountable for their actions, and to facilitate the interaction between citizens and their representatives to make the political process more effective”* (p. 19). Hence, this definition corresponds to the “maximalist” definition of Simon et al. (2017).

As to van Dijk (2013), the definition of digital democracy is linked to the use of digital media. He argued that digital democracy could be defined as *“the pursuit and the practice of democracy in whatever view using digital media in online and offline political communication”* (p. 4). He adds that *“the online-offline distinction should be added because political activities are not only happening on the Internet but also in physical meetings where mobile digital media are used for assistance”*. (p. 4)

Tsagarousianou (1999, in van Dijk, 2013), distinguished three claims in favor of digital democracy:

1. Digital democracy improves *political information retrieval and exchange* between governments, public administrations, representatives, political and community organizations and individual citizens.
2. Digital democracy supports *public debate, deliberation and community formation*.
3. Digital democracy enhances *participation in political decision-making* by citizens.

2. Typology for Digital Democracy

Simon et al. (2017) developed a typology to identify different aspects of digital democracy (see Figure 1). They state that it is grounded in practice so that it would be useful for practitioners, policymakers, and citizens in order to explain “*how democratic institutions can make use of digital technologies and how citizens themselves can be involved in the practice of everyday democracy – such as raising specific concerns, developing and scrutinizing legislative proposals, making decisions or holding public officials to account. The typology we set out below aims to set out these ‘everyday’ activities*” (p. 12).

Figure 1. A typology of digital democracy (Simon et al., 2017)



3. E-democracy and its Dimensions

Kneuer (2016) argued that e-democracy is a three-dimensional concept. It encompasses free and equal access to the Internet, e-participation, and e-government. These concepts are designed as a “concept tree” in the sense of Kneuer (2016). In what follows, the authors will define and present each of these concepts.

The first dimension called “free and equal access” which is also composed of two sub-dimensions: legal provision and technical infrastructure. These are considered as the basic conditions to allow e-

democracy. The former condition (legal provision) is considered as the ability of a country to guarantee Internet freedom use and content regulation for its citizens, public and private organizations. The latter is correlated to the former. In fact, the technical infrastructure is considered as a “*sine qua non*” condition to guarantee Internet freedom without “*any filtering or blocking by independent offices or regulatory bodies*” (Kneuer, 2016, p. 671).

van Dijk (2010, in van Dijk, 2013) states that the second dimension of e-democracy called “e-Participation” can be defined as “*the use of digital media to mediate and transform the relations of citizens to governments and to public administrations in the direction of more participation by citizens*” (p. 9). Likewise, he argued that e-participation is related and correlated to the concept of “the policy process” through its five phases, namely: agenda setting, policy preparation, decision making, policy execution, and policy evaluation.

The detractors of this new political approach consider that citizens can not intervene in all these five phases. The only ones are agenda setting, policy preparation, and policy evaluation. Applying e-Participation in decision making and policy execution is contested. In fact, decision-making, and policy execution are currently reserved for political representatives and public administrations executing the decisions of governments and parliaments. In the context of representative democracy, the only possibilities to intervene are e-voting and e-campaigning.

For Kneuer (2016), “e-participation” is considered as the second e-democracy dimension which is characterized by three basic elements. The first is based on the fact that e-Democracy necessitates a “*two-way channel communication*”:

- “a top-down direction – or in the words of Coleman and Blumler: e-democracy from above –
- and a bottom-up direction – or democracy from below (Coleman and Blumler, 2009: 90–139; see also OECD, 2003: 30)” (Kneuer, 2016, p. 671).

These two-way channel communications are considered by Kneuer (2016) as the two dimensions in e-participation.

The second element is “the decision-making process”. For Kneuer (2016), it is necessary to distinguish between “e-voting” and “engagement” in the decision-making process. In fact, there is a fundamental difference between these two notions which is based on the role given to the “citizen” in the political decision-making process.

In modern democracies one could distinguish between formal and informal ways to participate in decisions: “*Formal participatory modes are directed to the government, or the parliament; informal ones have no direct addressee, but should put pressure on the decision-takers through their broad public support of a specific claim. In the first case, they can be initiated both by a Citizens’ Initiative or by the government or parliament, and they can define specific groups that are entitled to participate, mostly relating to the group that is affected by the issue at hand. Examples of this are, decisions about an urban, or a regional issue, or decisions that concern specific interest groups, like trade unions, or Chambers of Commerce*” (Pohoryles, 2017, p. 9).

Pohoryles (2017) argued that the voting system is the only participative element in a classical form of democracy. He added that “... *the notion of “Stimme abgeben” (in English: to cast a vote), which literally means to delegate decisions to people, or parties and to refrain from any further influence on specific policies*” (p. 8).

Electronic voting or e-voting is defined as a “*type of vote which is done through electronic systems. (...) Electronic voting, also known as e-voting, implies the use of technology ... It is a form of eGov-*

ernance, but as eDemocracy as well. (...) With respect to eDemocracy one can expect, for instance, a higher participation rate and inclusion of the otherwise challenged - for being away from their domicile or for being physically impaired” (Pohoryles, 2017, p. 9).

The third element of the e-participation is e-monitoring. For Hindman (2009, in Kneuer, 2016), e-monitoring “is based on the assumption that through digital media citizens can not only gain better knowledge of grievances or undesirable developments but also have a stronger voice to express ‘alarm’” (p. 672).

The third dimension of e-democracy is “e-government” which is defined as “the use of ICT (...) as a tool to achieve better government’ (OECD, 2003: 11), in relation to the efficiency of public administration processes, public services for consumers/citizens, to contribute to specific policy outcomes (for example in health policy, environmental policy, and educational policy) or economic policy outcomes like reducing corruption, and to public management modernization” (Kneuer, 2016, p. 669).

III. Digital Participatory Platform (DPP)

1. Definitions

As specified above, “the practice of democracy using digital tools and technologies” is called “digital democracy” (Simon et al., 2017). However, among these digital tools, these authors distinguish the Digital Participatory Platform (DPP) as a specific kind of collaborative social media. In fact, Lee & Kwak (2012) distinguished between two groups of social media based on their main purpose (Kotler et al., 2010 in Lee & Kwak, 2012). The first group “is expressive social media, which enables people to express themselves by sharing with others text, picture, video, and music. Facebook, MySpace, Twitter, Youtube, Flickr, and Foursquare fall into this category of social media” (Lee & Kwak, 2012, p. 492). The second group is considered as a “collaborative social media, which enables people to work together to achieve common goals through interactive and social processes. Wiki and Google Docs are great examples of this type of social media” (Lee & Kwak, 2012, p. 492).

For Falco and Kleinhans (2018a) Digital Participatory Platforms (DPPs) are considered as a specific kind of collaborative social media. They “include all the features proper to ‘conventional’ social media (such as Facebook and Twitter): they are based on Web 2.0, allowing for user generated content, and sharing of such content. However, compared to conventional social media, DPPs also include different and more elaborate technological features ...” (p. 17).

2. Digital Participatory Platform (DPP): Potential and Implications

The majority of the researchers argues that we need more digital tools for actual collaboration and citizen engagement in local government activities (Desouza & Bhagwatwar, 2012; Ertiö, 2015; Williamson & Parolin, 2013; and Zavattaro & Sementelli, 2014 in Falco & Kleinhans, 2018). Indeed, this awareness and this trend were born following the call made by the US. President Obama in 2009 to increase the openness of the government which is based on three main principles: transparency, participation, and collaboration (Executive Office of the President, 2009 in Lee & Kwak, 2012). The open government finds its roots in the advent of the collaborative social media enabling the engagement and the participation of citizens in the government work. In fact, the social media “has changed the public’s expectations about how government work should be done ...” (Lee & Kwak, 2012, p. 492).

The implementation of a Digital Participatory Platform (DPP) takes into account the mandatory use of issues and challenges that local governments face in fostering online and offline citizen engagement (Falco and Kleinhans, 2018). These challenges and issues are seen as a prerequisite to identify actions and solutions for local governments. In fact, what any local government, especially municipality aspire or look for is not only implementing DPP but also how this later should be designed to foster citizen engagement and democracy. Therefore, social media enables the government to become a platform for public participation and collaboration (Lathrop & Ruma, 2010 in Lee & Kwak, 2012).

Wright (2016 in Ranchordás, 2017) claims that digital discussion forums, social media, electronic petition ('e-petition') websites, and several other technologies of participation have the merit of increasing citizen participation and improve the development of the participatory democracy to cope with the perceived crisis in a representative democracy. For Ranchordás (2017) professional digital platforms promote civic engagement and have a huge role in the promotion of citizen participation.

Finally, one can conclude that digital participatory platform has a real potential and a positive impact on participatory democracy and citizen engagement.

3. Digital Participatory Platform (DPP): a Key Factor in Developing Smart Cities

The study of the relationship between the DPPs and the development of smart cities could be explained by focusing on the role of citizen participation and e-participation in their development (Berntzen & Johannessen, 2016; Vrabie & Tîrziu, 2016). The authors stated that developing smart cities could be achieved through citizen participation and e-participation, which is based on the implementation of the DPPs.

For Berntzen and Johannessen (2016) Smart city is considered as "*a label or statement to indicate that the city is actively pursuing use of modern technology to increase the quality of life in urban space, both by improving the environmental conditions and delivering better services to its citizens*" (p. 2). However, a smart city concept could be considered more than a simple label. For Aurigi (2005 in López-Quiles and Rodríguez Bolívar, 2018) "*the smart city concept and other similar concepts such as intelligent cities, virtual cities, knowledge-based cities, digital cities or information cities have emerged on the idea that ICTs is central to the operation of the future city*" (p. 1).

For López-Quiles and Rodríguez Bolívar (2018), the implementation of ICTs in cities is a "*sine qua non*" condition to improve access and efficiency of public services through electronic administration and for improving e-participation as main elements to become smart. Moreover, Erkul (2014 in Vrabie and Tîrziu, 2016) argued that the four pillars of e-participation (greater government transparency, more focus on citizen needs, increased citizen involvement, improved government responsiveness) could lead to the creation and/or development of smart cities.

Thus, the DPPs could be considered as a key factor in developing smart cities (López-Quiles and Rodríguez Bolívar, 2018).

4. Digital Democracy Projects Around the World

This section aims to present some typologies of digital platform projects launched around the world. The authors focused on the empirical study of Falco and Kleinhans (2018), which is consistent with the main subject of this manuscript. There is another empirical study belonging to Simon et al. (2017), very interesting, however, their typology¹ does not correspond to the specificity of the current case study "Baladiaty".

Through a systematic review conducted by Falco and Kleinhans (2018), 113 active DPPs around the world have been identified, analyzed, and classified within a citizen-government relationship typology composed of five digital platforms: (1) *Consulting platforms*, (2) *Interaction platforms*, (3) *Reporting platforms*, (4) *Co-production platforms*, and (5) *Self-organization platforms* (of citizens).

The authors will focus only on three types of them. In fact, their features and objectives correspond to the “Baladiaty” case study. The first is called the *Consulting platforms* “... require citizens to provide their views, comments and preferences through consultations and surveys exist but usually sustain no options for the government to provide feedback, thus lacking an interactive and co-productive potential” (Falco and Kleinhans (2018, p. 9). Among the interaction platforms, Falco and Kleinhans (2018) distinguish “the reporting platforms that facilitate interaction between citizens and the local government on a specific practical issue to be solved (citizens report problems and the local government informs them when it has been fixed)” (p. 9).

Finally, Falco and Kleinhans (2018) found that the “Co-production platforms are also quite widespread with many examples from around the world (e.g. *Block by Block*, *Commonplace*, *coUrbanize* ...” (p. 9). The “co-production” means that the public sector and citizens making better use of each other’s assets and resources to achieve better outcomes and improved efficiency.

BALADIATY: A PILOT DIGITAL PARTICIPATORY PLATFORM IN TUNISIA

The presentation of the different stages followed in order to implement a pilot digital participatory platform (DPP) called “Baladiaty” in Tunisia, is considered as the focal point of this paper. “Baladiaty” is a DPP designed, developed and commercialized by ARSELA in 2019, a Tunisian start-up located in the governorate of Sousse.

1. The Context: The Participatory Democracy in Tunisia: Legal Framework and Perspectives

It is important to note that participatory democracy was anchored in the 2014 constitution² in Tunisia by several articles (Article 31 and Article 32) which address the notion of decentralization and citizen participation on the basis of which an organic law of the code of collectivities was adopted by the assembly of representatives of the Tunisian people (ARP).

Chapter VII of the Tunisian constitution of 2014 was devoted to local power and includes Articles 131, 139 and 140 which were the most explicit on local governance and citizen participation which concerns the partnership between ARSELA and the municipalities.

These articles mark the transitional approach towards decentralization and the strengthening of the process of evolution towards local decision-making to be reinforced by methodological and operational tools to implement. In this sense, the Organic Law of Local Authorities of May 2018 defined, inter alia, the role of civil society as a stakeholder in decision-making at the local level and at different levels of participation and at different stages of municipal works which will now be governed by elected municipal councils. After the first municipal elections of 2018, the new municipal councils moved from the election period to the executive term of municipal works, in order to fulfill the duty for which they had been elected and to satisfy the electors to whom they committed themselves.

In Tunisia, the organic law n ° 2018-29 of May 9, 2018, relating to the code of local authorities, dedicates for the first time the existence of a local decision-making referendum which can be initiated by local authorities or by 1/10 of the community residents participating in community programs and projects. This law also regulates the contribution of civil society to the development and monitoring of local development projects and provides a legal framework for the consultation process for the preparation of development plans. This code is based on the foundations of Article 139 of the 2014 Constitution mentioned above.

The system of local democracy is now hybrid between two types of democracies: representative and participatory. Representative by elected municipal councils that embody a representative democracy based on the election by universal suffrage of representatives of the municipality or local government deliberative bodies in the case of regional and sectorial councils.

Participatory through the involvement of civil society and citizens in decision-making at different levels (information, consultation, and co-decision) and degrees (organizations or citizens) depending on the nature of the projects (administrative, structuring and local). In general, participatory democracy covers different modes of intervention of citizens, individually or through associations, in the elaboration of local decisions.

2. Baladiaty: The Main Features

“Baladiaty” is a technically reliable and ready-to-use solution. It would enable municipalities to join the ranks of smart and innovative cities and move from a fragmented environment to a global and integrated management while simplifying operations and improving services to citizens.

The implementation of the proposed solution is part of a more global approach to digitizing the services provided to citizens to facilitate their relationship with the administration. In fact, the proposed digital solution will make it possible, on the one hand, to improve and simplify access to municipal services by citizens in relation to existing administrative procedures and to facilitate the missions of municipal officials on the other hand.

The “Baladiaty” platform was designed to gather all the tools necessary for effective communication between the municipality and the citizens. It allows to inform, to alert, to exchange, to collect the opinions of the citizens and much more. “Baladiaty” would be an ideal tool to promote good local governance and citizen participation.

In the new context of elected municipal councils, since participatory democracy has been constitutionally imposed since 2014 and supported by the Local Authorities Code of 2018, each municipality wants residents to be constructively involved in local development. By reporting incidents or proposing improvements, the responsible citizen is valued when he is warned that a particular report has been corrected, that his requests have been taken into account.

The number of municipalities in Tunisia is 350. In the context of decentralization, municipalities have free management tools, independent of the ministry, which gives them the possibility to choose management systems specific to their needs as Baladiaty product. The platform has been tagged with the National Authority for Protection of Personal Data to be credible but also optimizes the exploitation of its inputs and their commercialization in the rules of law.

Thus, “Baladiaty” has three strategic pillars of the public value management (PVM) also called the Moore (1995)’s strategic public value triangle³: 1) the definition of the public value 2) the legitimacy and support 3) the operational capabilities.

- Definition of the public value:

Strategic objectives:

- Build a constructive ecosystem and enhance the communication between the municipality and citizens.
- Projects matching with citizens' needs and priorities.
- Agile administrations.
- Attract more stakeholders.
- Make proof and enforce participative democracy and governance.

Strategic results:

- Enhance citizen engagement with their municipalities of at least 60% by 2022 and 80% by 2025.
- Increase citizen's satisfaction towards municipality services and projects of 50% by 2022 and 80% by 2022.
- Boost the trust between the 2 parties 4/5 citizens trust their municipality.
- Decreasing the number of social and solidarity projects/ sustainability projects.
- Bring closer the municipality's stockholders (associations, sponsors, project owners, etc.).
- Becoming a real modern and smart City by 2025.
- **Legitimacy and support**
 - The Tunisian constitution and the laws foster and support participatory democracy.
 - Local community law that pushes municipalities to find projects to empower their own governance.
 - "Baladiaty" is aligned with the government vision of digitalizing the administration services.
 - "Baladiaty" is aligned municipalities mission to improve their services, impact, and communication with citizens.
- **Operational capabilities**
 - "Baladiaty" enables municipalities to join the ranks of e and innovative cities and move from a fragmented environment to holistic and integrated management, by simplifying operations and improving services to citizens.
 - "Baladiaty" will improve and simplify access to municipal services by citizens in relation to the existing administrative procedures and will facilitate the missions of municipal officials.

To determine the value of the platform, the authors will take a closer look at its components. Specifically, the platform has the following three functionalities: (1) Complaints and petitions service; (2) Active Citizenship Space - Survey; and (3) Space of Participatory Democracy - Proposal.

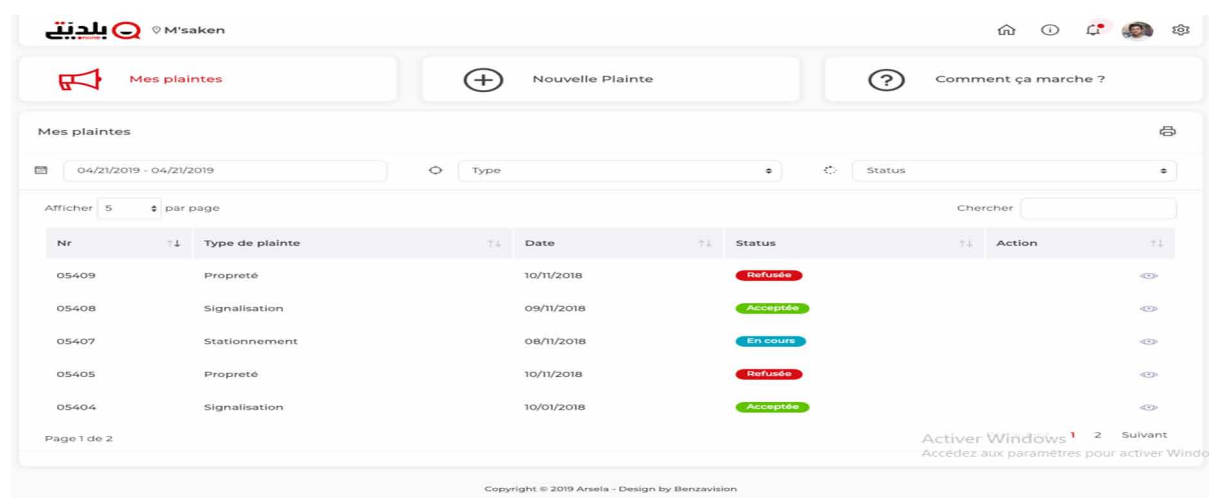
Complaints and Petitions Service

A complaint is the expression of a citizen's dissatisfaction with municipal services. A complaint may also consist of a broken ferry, a defective lamppost, an act of vandalism, it is in fact a malfunction signal within the jurisdiction of the municipality, allowing the municipal team or technical services of the municipality to intervene quickly when an anomaly is encountered on public roads.

The ability to trace geo-located information in a few seconds to the municipality agents represents a considerable asset in the management of a commune, promotes participatory democracy and mobilizes the citizens who can report any dysfunction or problem encountered in the city: pothole, rubbish, lighting problems, bulky objects, abandoned vehicles, defective traffic lights, etc.

The complaints and petitions module should allow citizens to file petitions and complaints (see Figure 2). The solution consists in proposing a form to identify the applicant who is led to make the choice of the type of complaint as well as the municipal service concerned, the seizure of the description of the complaint, to join a proof (photographs, documentation, etc.) and even share his map position.

Figure 2. Baladiaty User Interface - Complaints Service



After receiving a request and complaint form specifying the subject of the request, the municipality is responsible for processing the request and responding to it within 3 to 21 days. The applicant is notified when processing his request through a monitoring dashboard allowing him to access the information related to his request in real-time. This information consists of locating the progress of the processing of the complaint. The solution proposes to provide the staff of the municipality visualization of the report of complaints filed with the municipality on a dynamic map.

Active Citizenship Space – Survey

The solution proposes to provide a space where municipalities can question the population on subjects of their choice and can share the achievements and advancements of their communal projects (see Figure 3).

Before setting up a project for your municipality, it is important to survey the local population to take public opinion into account: how will citizens react? ; What are their expectations, their wishes, and their preferences?

In addition, it could be a future reference database available to the municipality.

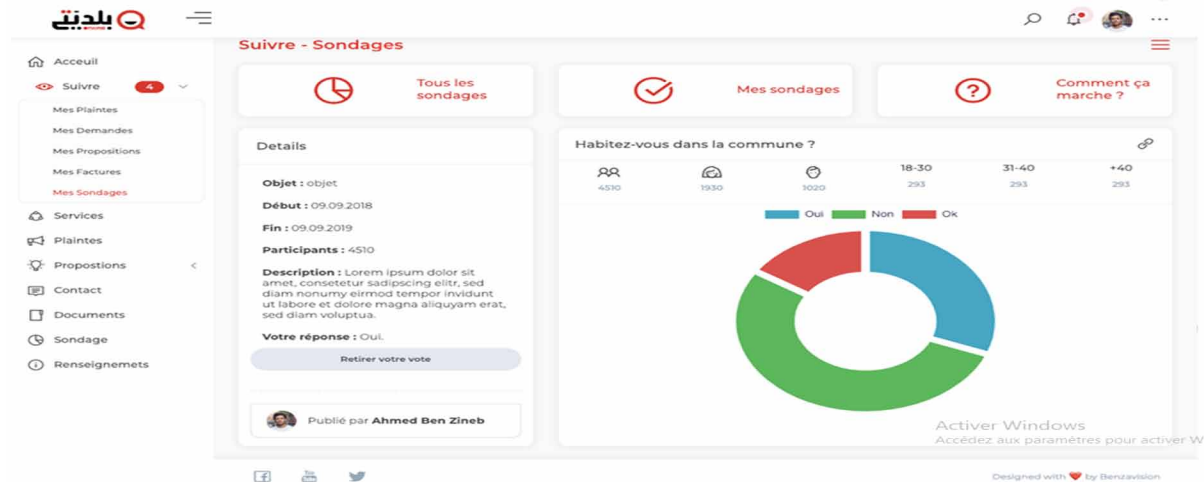
Space of Participatory Democracy – Proposal

Giving citizens the opportunity to suggest an idea to their community allows not only to solve everyday problems but also to make decisions by sharing decisions by bringing leaders closer to the governed.

The advent of new technologies makes it possible to transpose this tool of participatory democracy to digital, and thus allows any citizen user to make a suggestion to his municipality, from his computer, tablet or mobile phone.

Figure 3. Baladiaty Administrator Interface - Active Citizenship Space – Survey

12



The “Suggestion” module should allow citizens to propose project suggestions. It will be generated via the citizen profile of the user. The module will display the number of user evaluations of the platform (commune and citizens) of the shared suggestion.

3. Launch of “Baladiaty”

The platform was adopted by the municipality of Msaken, it was delivered on July 12, 2019 with its user guide and launched online for members of the city council.

The official public launch is scheduled for 17 August 2019 in Msaken during a press conference followed by a communication campaign with citizens who will be invited to download the application and use it. Communication materials have been developed by ARSELA: leaflets, brochures, roll-ups, banners, and billboards. Volunteers will form a door-to-door action that will last three days throughout the town of Msaken. A communication campaign has already been launched on social networks, specifically on Facebook.

The development of a communication strategy dedicated to the launch of the Baladiaty in Msaken helps to set up appropriate communication, identify objectives, clarify the challenges, decide on the reputation and brand strategy that projects’ designers aspire to communicate to the stakeholders of the project and implement and affirm the vision and objectives of ARSELA.

The strategic objectives of this communication strategy are:

- To inform of the existence of an alternative of communication with the commune.
- Trigger the curiosity of the potential user: Give the citizen the desire to discover and browse the features of the platform.
- Creation of a brand image for “Baladiaty” in Msaken.
- Communicate around an image of exclusivity, modernity, and credibility with the aim of consolidating trust and the close relationship between the citizen and his municipality.
- Act: Account creation & use/ownership of the platform by the citizen.

In addition, the project was presented to the Tunisian Confederation of Mayors and the National Federation of Tunisian Cities. ARSELA team obtained an agreement in principle of collaboration that will start by providing the Baladiaty platform to ten municipalities in order to reap the benefits and will allow ARSELA to adjust the solution, promote it nationally and propel it to the MENA (the Middle East and North Africa) region.

4. Baladiaty: The Main Challenges

As a nascent digital participatory platform, and based on an empirical literature review, the authors will present a set of potential challenges that “Baladiaty” will likely face.

Simon et al. (2017) provided a set of challenges faced by three digital participatory platforms implemented by municipal governments. These three case studies, are Decide Madrid (Spain), Better Reykjavik and Better Neighbourhoods (Iceland), and Madame Mayor, I have an idea (France).

The main challenges to overcome are:

- For Decide Madrid (Spain):
 - Financial and human resource requirements,
 - Improving citizen visibility of proposals,
 - Broadening participation and increasing representativeness (Simon et al., 2017, p. 51).
- For Better Reykjavik and Better Neighbourhoods (Iceland):
 - Ensuring breadth of participation and representation,
 - Increasing motivations for participation,
 - Lack of public understanding of the platform and process,
 - Managing resource requirements (Simon et al., 2017, p. 46).
- For Madame Mayor, I have an idea (France):
 - Processing and developing ideas,
 - Finding the right balance of information provision,
 - Generating ideas to address more complex issues,
 - Raising awareness and engagement (Simon et al., 2017, p. 55).

FUTURE RESEARCH DIRECTIONS

Although “Baladiaty” was designed taking into consideration the main features of the best DPPs in the world, the authors are aware that the role of local governments in the success of this pilot project is crucial and strategic. In fact, Spirakis et al. (2010) argued “..., the effect of e-government depends not only on the development of technology but also on the organizational resources and the strategic perspectives” (p. 77). They also argued, “The implementation of ICTs in the public sector can be used as a strategic tool of growth and also as an answer in the current challenges of globalization” (p. 78). Moreover, Falco and Kleinhans (2018) stated “Despite a growing number of web-based and mobile-based platforms that enable information sharing and interaction between government and citizens, scholars have highlighted that use of DPPs is not yet interactive and is not being able to create two-way communication (Williamson & Parolin, 2013; Ertio, 2015)” (p. 4).

Therefore, the main question to answer is how local government will overcome these challenges if they exist. In addition, do local governments face specific challenges? How to measure the impact

of “Baladiaty” on local participatory democracy and citizen engagement in a nascent democracy like Tunisia? What role do technology and networks play in governance? How do they affect traditional structures and organizations? What about public value? What is the nature of the relationship between crowdsourcing and digital participation as a pillar of digital democracy? Etc.

CONCLUSION

This paper belonging to the topic of electronic government aimed to present the implementation of a pilot digital participatory platform (DPP) called “Baladiaty” in order to foster local government-citizens collaboration and participation in Tunisia. However, it is obvious to note that the description of the single Tunisian experience (“Baladiaty”) cannot produce a valuable generalization concerning the role of e-democracy, digital participatory platforms and their impact on collaboration between local government and citizens.

Although this article presented a pilot digital participatory platform in a nascent democracy (Tunisia), it could be considered as a very rich theoretical manuscript in which the authors introduced and defined various basic concepts in the field of public management, namely: democracy, e-democracy, representative democracy, participatory democracy, e-participation, e-government, DPPs, etc.

As explained all over this manuscript, the implementation of a digital participatory platform has a critical and strategic perspective not only for governments or for local governments (e.g. municipalities), but also for policymakers who will improve the sense, practice, and meaning of democracy for all the citizens. Hence, all the concepts described here have a strategic role in the relationship between new participatory models and their impact on collaboration between local government and citizens.

REFERENCES

- Abu-Shanab, E. A. (2015). E-democracy: The fruit of e-government. *International Journal of Technology and Globalisation*, 8(1), 15–28. doi:10.1504/IJTG.2015.077873
- Adejuwon, K. D. (2012). From E-government to E-governance: Wither African Public Administration? *Advances in Arts, Social Sciences and Education Research*, 2(1), 63–75.
- Berntzen, L., & Johannessen, M. R. (2016). The Role of Citizen Participation in Municipal Smart City Projects: Lessons Learned from Norway. In *Smarter as the New Urban Agenda* (pp. 299–314). Springer International Publishing. doi:10.1007/978-3-319-17620-8_16
- Bratton, M., Mattes, R., & Gyimah-Boadi, E. (2005). *Public Opinion, Democracy, and Market Reform in Africa*. Cambridge University Press.
- Clive, C. (2012). Towards understanding eParticipation in the public sphere. *Review of Business Research*, 12(1), 140–146.
- Dalton, R. J., Shin, D. C., & Jou, W. (2007). Understanding Democracy: Data from Unlikely Places. *Journal of Democracy*, 18(4), 142–156.

- Falco, E., & Kleinhans, R. (2018a). Beyond technology: Identifying local government challenges for using digital platforms for citizen engagement. *International Journal of Information Management*, 40, 17–20. doi:10.1016/j.ijinfomgt.2018.01.007
- Falco, E., & Kleinhans, R. (2018b). Digital Participatory Platforms for Co-Production in Urban Development: A Systematic Review. *International Journal of E-Planning Research*, 7(3), 1–27. doi:10.4018/IJEPR.2018070105
- Floridia, A. (2013). *Participatory Democracy versus Deliberative Democracy: Elements for a Possible Theoretical Genealogy. Two Histories, Some Intersections*. In 7th ECPR General Conference, Bordeaux, France.
- Gil de Zúñiga, H., Veenstra, A., Vraga, E., & Shah, D. (2010). Digital Democracy: Reimagining Pathways to Political Participation. *Journal of Information Technology & Politics*, 7(1), 36–51. doi:10.1080/19331680903316742
- Grewal, S. (2019, February). *Tunisian democracy at a crossroads*. Retrieved from <https://www.brookings.edu/research/tunisian-democracy-at-a-crossroads/>
- Kneuer, M. (2016). E-democracy: A new challenge for measuring democracy. *International Political Science Review*, 37(5), 666–678. doi:10.1177/0192512116657677
- Lee, G., & Kwak, Y. (2012). An open government maturity model for social media-based public engagement. *Government Information Quarterly*, 29(4), 492–503. doi:10.1016/j.giq.2012.06.001
- Lévesque, B. (2014). La nouvelle valeur publique, une alternative à la nouvelle gestion publique? *Revue vie économique*, 4(2), 1-18.
- López-Quiles, J. M., & Rodríguez Bolívar, M. P. (2018). Smart Technologies for Smart Governments: A Review of Technological Tools in Smart Cities. In M. Rodríguez Bolívar (Ed.), *Smart Technologies for Smart Governments. Public Administration and Information Technology* (Vol. 24). Springer. doi:10.1007/978-3-319-58577-2_1
- Lundström, M. (2004). *Three Concepts of Democracy in Swedish Political Science*. ECPR 32nd Joint Sessions of Workshops, Workshop 12: National Traditions of Democratic Thought, Uppsala, Sweden.
- Pohoryles, R. J. (2017). eDemocracy, opportunities and challenges - The concept of modern democracy and the use of ICTs: Definitions. In European Liberal Forum. In *eDemocracy and eParticipation The precious first steps and the way forward* (pp. 7-11). The European Liberal Forum asbl. with the support of the Project office for Southeast Europe of the Friedrich Naumann Foundation for Freedom. Co-funded by the European Parliament.
- Ranchordás, S. (2017). Digital agoras: Democratic legitimacy, online participation and the case of Uber-petitions. *The Theory and Practice of Legislation*, 5(1), 31–54. doi:10.1080/20508840.2017.1279431
- Simon, J., Bass, T., Boelman, V., & Mulgan, G. (2017). *Digital Democracy: The tools transforming political engagement*. Nesta.
- Spirakis, G., Spiraki, C., & Nikolopoulos, K. (2010). The impact of electronic government on democracy: E-democracy through e-participation, *Electronic Government. International Journal (Toronto, Ont.)*, 7(1), 75–88.

van Dijk, J. A. G. M. (2013). In I. Snellen & W. van de Donk (Eds.), *Digital Democracy: Vision and Reality. Innovation and the Public Sector* (pp. 2–3). IOS – Press.

Vrabie, C., & Tîrziu, A. M. (2016). E-participation – a Key Factor in Developing Smart Cities. *EIRP Proceedings, 11*, 2010-2015.

ADDITIONAL READING

Burkhardt, D., Zilke, J. R., Nazemi, K., Kohlhammer, J., & Kuijper, A. (2014). Fundamental Aspects for E-Government. In P. Sonntagbauer, K. Nazemi, S. Sonntagbauer, G. Prister, & D. Burkhardt (Eds.), *Handbook of research on Advanced ICT integration for Governance and Policy Modeling* (pp. 1–18). IGI Global. doi:10.4018/978-1-4666-6236-0.ch001

Desouza, K. C., & Bhagwatwar, A. (2014). Technology-Enabled Participatory Platforms for Civic Engagement: The Case of U.S. Cities. *Journal of Urban Technology, 21*(4), 25–50. doi:10.1080/10630732.2014.954898

Evans-Cowley, J. S., & Kubinski, B. (2015). There's an App for That: Mobile Applications That Advance Urban Planning. In C. Nunes Silva (Ed.), *Emerging Issues, Challenges, and Opportunities in Urban E-Planning* (pp. 33–45). IGI Global. doi:10.4018/978-1-4666-8150-7.ch002

Gordon, E., & Mihailidis, P. (2016). *Civic Media: Technology, Design, Practice*. The MIT Press. doi:10.7551/mitpress/9970.001.0001

Kleinhans, R., Van Ham, M., & Evans-Cowley, J. (2015). Using social media and mobile technologies to foster engagement and self-organization in participatory urban planning and neighbourhood governance. *Planning Practice and Research, 30*(3), 237–247. doi:10.1080/02697459.2015.1051320

KEY TERMS AND DEFINITIONS

DDP: A kind of social media. It is considered as a collaborative social media.

Direct Democracy: A form of democracy in which citizens play a direct role in political decision-making without intermediaries.

E-Democracy: Also called digital democracy. It is defined as the practice of democracy by using information and communication technologies.

E-Government: It is defined as the use of digital tools for exchanging information, providing services and transacting with citizens, businesses, and the rest of the stakeholders.

E-Participation: Is one of the e-democracy dimensions. It is related to e-government and e-governance. It is about fostering citizen's engagement.

Participatory Democracy: It is defined as the involvement of citizens in decision-making on public policy.

Representative Democracy: Also called indirect democracy. It is the contrary of the participatory democracy. People elect officials or government representatives to handle legislation and ruling the country on their behalf.

ENDNOTES

- ¹ For more details, please refer to Simon et al. (2017).
- ² https://www.constituteproject.org/constitution/Tunisia_2014.pdf
- ³ “Moore (1995)’s strategic public value triangle is based on three concerns that must be faced by political leaders and public service managers. First of all, the definition of public value, concretely clarifying the objectives and purposes of the services that must include the production of value not only for the individuals directly affected but also for the communities concerned, which implies making the link between the user and the citizen and to appeal to the deliberation since the public value is plural and that it is never defined once and for all. (...) Secondly, the concern for the public value cannot be realized without an authorized and legitimate environment since it is necessary at the same time the support of the public authorities and the construction of a coalition of the stakeholders (users, communities, private sector, public sector, voluntary sector) with diversified interests. Thirdly, the operational capacity to build goods and services is required, which refers to the feasibility in terms of mobilizing operational resources (financial, personal, skills, technology) inside and outside.” (Lévesque, 2014, p. 9).

Local Government, Decentralisation, Devolution, and Service Delivery in Zimbabwe

12

Tawanda Zinyama

University of Zimbabwe, Zimbabwe

INTRODUCTION

Zimbabwe is one of only a handful of countries in sub-Saharan Africa that have taken part in the wave of decentralization and devolution. The process differs from the efforts of the 1980s, 1990s and 2000s in that this time the government of Zimbabwe seems ready to relinquish genuine power to the subordinate levels. Devolution of power is enshrined in Zimbabwe's 2013 Constitution as one of the country's founding values and principles. However, as was the case with in the 2000s, it is sometimes easy to romanticize about the benefits of devolution and fail to take into account the detailed work required before it benefits people at the local level. The devolved administrations exercise considerable autonomy when it comes to tackling poverty: for instance, major areas of economic development, education, health and housing policy are the responsibility of the devolved institutions.

This paper seeks to address the following questions: in what ways can devolution contribute to socio-economic and greater welfare of the people? What are the principles guiding decentralization and devolution? What are the ideal institutional frameworks for Provincial and Metropolitan Councils? How are these Councils going to raise and collect revenue? Will the devolution address the regional injustices and marginalization? Is there shared understanding of devolution in Zimbabwe? The paper is organized into five sections. First section provides the background and conceptual framework. Section two presents Zimbabwe's local government system focusing on the rationale, objectives and principles of devolution, innovations of the local government and planning as an instrument of social mobilization. Section three focuses on constraints on local governance such as lack of citizen consultations, unitary state, slow pace of legislation alignment, center-local relationships and violation of legislation among others. The fourth section covers solutions and recommendations such as assistance and support, public finance management system, politicization of local governance in Zimbabwe, civic culture creation and multi-stakeholder participation. The fifth section presents the future research directions. The last section provides a conclusion.

BACKGROUND AND CONCEPTUAL FRAMEWORK

Decentralization and Devolution

Interest group politics suggest that politicians at the center have little wish to cede their vast powers, notably those over public finances, to the local level. They embark on meaningful local reforms, including decentralization and devolution, only when they sense authentic benefits to themselves and their

DOI: 10.4018/978-1-7998-3479-3.ch109

political supporters from incorporating the local government into the mainstream political process, for example, when this gives them a political advantage over their competitors at the center or when, to generate welfare-enhancing growth, create broader markets (Zinyama, 2019; Chigwata, 2018) for urban-based industries or support the agricultural activities of the elite, it becomes necessary to liberalize the economy as well as domestic politics.

Decentralization is a generic term for the diffusion of governmental authority and power away from the national center to other institutions at other levels of government or levels of administration (Moyo and Ncube, 2014). Devolution is generally classified as the most extensive form of decentralization that diffuses substantial governmental powers, authority, responsibilities and resources to local units. Such units exercise a measure of autonomy, which is “the extent to which local governments have discretion in carrying out their duties and obligations” (Masunungure and Ndoma, 2013; Chigwata, 2019). However, such autonomy cannot be equated to the independence enjoyed by a sovereign State, given that local units exercising devolved powers are not at par with the central government. Unlike federal systems where the regions or provinces maintain their own independent power and cede or receive some authority to or from the federal government, devolved units do not occupy the same sort of position in relation to the central government (Chigwata, 2019). The dimensions of devolution include centralization, administrative, fiscal, political, delegation and deconcentration (Barnett, et al. 1977; Poverty Reduction Forum Trust, 2019). Devolution aims to create local governments that are independent, autonomous and usually have exclusive authority over explicitly defined functions. Political autonomy is meaningless if it is not accompanied by fiscal autonomy, which entails the ability to raise and spend revenue (Moyo, 2013; Chigwata, 2018).

ZIMBABWE’S LOCAL GOVERNMENT SYSTEM

Local government is recognized in Zimbabwe as an important tier or sphere of government closet to the people, capable of providing a participatory and accountable style of governance for local communities (Zinyama, 2019, Chigwata, 2018; Moyo, 2013). It helps to provide service provision to communities in an equitable and sustainable manner. Chapter 14 of the Zimbabwe Constitution (2013) provides the framework for devolution of governmental powers and responsibilities in Section 264. Likewise, in Section 5 of the Constitution, the tiers of government provided are the national government; provincial and metropolitan councils; and local authorities. Since the adoption of the 2013 Constitution, the government has not yet crafted legislation to devolve power to Provincial and Metropolitan Councils.

Sections 268 and 269 of the Constitution of Zimbabwe establish Provincial Councils and Metropolitan Councils respectively. The Provincial and Metropolitan Councils are a middle and critical link between government and citizens.

Section 270 of the Constitution requires Provincial and Metropolitan Councils to undertake social and economic development in their respective provinces, including:

- planning and implementing social and economic development activities in its province;
- coordinating and implementing governmental programs in its province;
- planning and implementing measures for the conservation, improvement and management of natural resources in its province;
- promoting tourism in its province, and developing facilities for that purpose;
- monitoring and evaluating the use of resources in its province; and
- exercising any other functions, including legislative functions, that may be conferred or imposed on it by or under an Act of Parliament.

Devolution Rationale

A number of factors have contributed to the highly centralized state structure in Zimbabwe though having a good track record of representative democracy. Firstly, colonialism bequeathed a historical legacy of centralized government. Secondly, social and economic crises since late 1990s have been powerful stimuli for centralization. Thirdly, powerful political and bureaucratic vested interests have consciously thwarted attempts to fully decentralize and devolve power. Fourthly, successful decentralization requires certain pre-conditions (Zinyama, 2019; Chigwata, 2018; Moyo, 2013) such as an appropriate legal and administrative framework, a local information base, capacity-building programs and civic culture. To make the fulfilment of these pre-conditions a requirement for introducing democratic decentralization can become an argument for postponing devolution of powers to the lower levels.

Devolution is, therefore, seen as a necessary vehicle for doing away with the over-centralized system of government, deepening democracy, promoting locally driven development, improving the delivery of public services, and promoting national integration and peace while recognizing diversity. However, the concept of devolution is very contentious, emotive and divisive with strong regional overtones in Zimbabwe (Masunungure and Ndoma, 2013). It is also a frequently misunderstood and sometimes deliberately distorted term (Ibid.). For example, former President Robert Mugabe thought devolution would divide the country into small pieces and cause disunity. His argument was that:

Those things are done in big countries, not a small country like ours.... Some are talking about separating Matabeleland region to become a country, that is impossible we don't want that (Newsday, 14 March 2012).

However, the late then Prime Minister, Morgan Tsvangirai disagreed with Mugabe:

Devolution does not mean secession. It does not mean separation; it is not about tribalism. It is about sharing the national cake equally (Masunungure and Ndoma, 2013).

What then has changed to justify the current optimism regarding the developmental impact of decentralization and devolution? Three factors can be noted in this regard. First, decentralization, at least in theory, holds great potential for development. It can be said to possess a dual mandate. Decentralized government provides space for people to participate in the formulation of policies that affect them directly, including the setting of local taxes, provision of social services and ensuring security of life and property (Zinyama, 2019). It is also assumed that it is easier to exercise inclusive politics, which enable communities and groups to influence policies that affect their daily lives, at the local than at the national level. Indeed, for many aid agencies, local participation has become the focus of their aid programs. Besides, decentralization and devolution are assumed to improve decision-making at the local level and to raise the quality of governance. Also, by encouraging local participation, decentralization and devolution enhance resource mobilisation, including development aid and ensure a more efficient use of resources. This increases the production of goods and services and, thus, local welfare (Bossuyt and Gould, 2000).

Second, the growing rural-urban divide in Zimbabwe and the undemocratic provincial structures have condemned the majority of the masses to unending poverty and wretchedness – without basic education, healthcare, shelter and nutrition – the alternative being migration to the squalor of city slums. The current power pyramid needs to be reversed. Everything that can be decided and implemented optimally at the local level should be kept at that level leaving the rest to be taken at national level. It is also necessary to move beyond the confines of representative democracy. Forums must be created so that the ordinary

citizens can directly intervene, deliberate and decide in the governance and development processes. Democracy, from being a ritual of periodic elections, needs to be extended to grassroots level.

Third, democratic decentralization and devolution are an important political response to Zimbabwe's development crisis. The expectation is that devolution accelerates socio-economic growth and creates a new model of growth with equity. Democratic decentralization can contribute to improvements in efficiency of implementation, particularly if the development process is made participatory and transparent. This helps prevent misuse of resources and allows for better monitoring of programs. Participation can also help in tapping dormant local resources in the form of monetary donations, material contributions and voluntary labor (Zinyama, 2019). In labor-surplus communities with disguised unemployment, community participation can mobilize significant contributions of human resources to create social and physical infrastructure (Ibid.).

The Objectives and Principles of Devolution

Section 264(2) of the Constitution provides the objectives of devolution. Devolution is necessary to “give powers of local governance to the people and enhance their participation in the exercise of the powers of the State and in making decisions affecting them” (Government of Zimbabwe, 2013). It seeks to promote democratic, effective, transparent, accountable and coherent government. Given that Zimbabwe is an ethnically diverse country, devolution is important to preserve and foster the peace, national unity and indivisibility of Zimbabwe. While there is nothing wrong with nationally led-development with respect to ‘national issues’, devolution is a means of “recognizing the right of communities to manage their own affairs and to further their own development” (Ibid.). If properly designed, devolution can ensure the equitable sharing of local and national resources. For locally driven development to have a higher chance of success, devolution is necessary to transfer responsibilities and resources from the national government in order to establish a sound financial base for each provincial and metropolitan council and local authority.

The objectives of devolution are supported by the general principles of provincial and local government enshrined in section 265 of the Constitution. The Constitution requires these local governments to exercise their functions in a manner that does not encroach on the geographical, functional or institutional integrity of another tier of government. They are encouraged to “cooperate with one another, in particular by (i) informing one another of, and consulting one another on, matters of common interest (ii) harmonizing and coordinating their activities” (Ibid.). While they are not hard rules, these principles demand positive action from provincial and local governments.

These functions are vague and ambiguous. They are a mere framework and understandably so because the crafters of the Constitution would not have expected to specify the functions in the Constitution for purposes of brevity. There is need to specify how central government functions such as education, health, roads, electricity provision, land-use management, licensing, transport and registry among others are to be shared between central government and Provincial Councils.

However, the Constitution does not define what socio-economic development entails. It is not clear whether this includes the power to adopt policies and spend budgets on matters that relate to socio-economic development. In general, socio-economic development encompasses almost everything, from social related services, such as, the provision of health, education and sanitation, to economic orientated activities including tourism, natural resources extraction and job creation. Legislation and policies that clarify the parameters of the socio-economic development function of the councils, among other things, is required.

Innovations of the Local Government in Zimbabwe

The Zimbabwean government sees the devolution of power as a key means of introducing popular democracy and fostering local governance. This promotes capacity building at the local level and ensures that there are local inputs in the decision-making process, especially with respect to service delivery. These factors would in turn help foster a local sense of ownership of development programs. The devolution is enshrined in the Zimbabwean Constitution of 2013, with amendments made to the Provincial Councils and Administration Act [Chapter 29:11] which spells out the mechanisms of decentralization and devolution. The amendments have been prompted by the need to align the current Provincial Councils and Administration Act [Chapter 29:11] with the provisions of the Constitution of Zimbabwe which advocates for the devolution of power to Provincial and Metropolitan Councils.

Devolution in Zimbabwe: Goals, Policy Innovations and Outcomes

What impact would these far-reaching interventions would have on local government's ability to deliver services and to what extent would local populations be involved in the running of their affairs? The main goals of decentralization and devolution process in Zimbabwe, along with policy innovations intended to achieve results are as follows.

Table 1.

Goals	Policy Innovations	Comments on Outcomes
Involvement of local people in the management of their own affairs	Identify local problems and find solutions to be implemented by locally elected organs	Raising local people's interest in the management of their affairs is probably the strongest outcome of devolution effort in Zimbabwe
Democratization of the decision-making process	Elect local leaders instead of having them appointed by central government Planning as an instrument of social mobilization	Local level democracy is taking longer to take hold, probably for lack of precedent. Equitable socio-economic development
Mobilisation of local resources	Make budgets and prioritise expenditure according to needs at the local level	Few local governments are financially independent from the center and they continue to be beholden to government
Reduction in costs for service delivery	Improve social infrastructure and expand social service provision in the countryside	Scarcity of resources makes the undertaking of extensive infrastructure and service extensions difficult Poverty reduction and improved living standards
Improved efficiency and accountability at the local level	Enhance human resources by appointing professional staff to run local services.	The quality of staff recruitment at the local is improving though lack of resources implies that many are not filled. Budgets should be used as management tools. The central government uses the budgets to monitor and evaluate the budgetary and financial management performance of provincial councils. Streamlined budgets facilitate frequent audits of financial statements.
Peace Promotion and Conflict Resolution	Establishment of National Peace and Reconciliation Commission Outreach programs for publicity	Territorial justice Development and democracy National integration and peace

Brief discussions of these goals follow.

The empowerment of local populations by involving them in the identification of their problems and the finding of solutions should be the driving feature of devolution in Zimbabwe. Local councils are composed of elected local people, who often hold personal stakes in the welfare of the province. Such individuals are expected to have considerable interest in the deliberations of the councils and to defend the interests of their constituencies against, for example, the demands of the central government. Some regions such as Matabeleland, Masvingo and Manicaland may start feeling that they are getting a fair deal in terms of access to resources. Furthermore, the devolution program, by involving hundreds of thousands of people, hitherto totally ignored, in the political process at the local, may stimulate a rapid growth in social responsibility and political creativity. This may increase the involvement of women and youth in local politics.

Democratization of decision-making: through local councils, the powers and responsibilities for making decisions over issues of local concern in the province would be democratized. This represents a fundamental shift from the traditional procedure where policies implemented at the local level were made by state bureaucrats based in the capital city and, thus, far away from the people affected by those decisions. This will help communities reach consensus on political and development issues. This is because the devolution system moves away from the unbridled and comprehensive subordination of locality to center of the past to a relationship steeped in consensus building, participation in policy formulation and negotiation. Increased transparency in local government would improve the stakeholders' capacity to expose the practices of malfeasance and corruption, which in the past often went undetected at the central level and to initiate corrective action.

Mobilization of local resources: proximity to local policy-makers would increase tax compliance, because the population not only would be able to see the direct benefits, in terms of improved services, but also would be able to monitor the use of resources and, via the provincial councils, hold errant officials accountable for abuse of office.

Improvement of service delivery: resource mobilization is closely related to the local government's capacity to improve service delivery. Local councils have assumed a number of functions that were previously performed by line ministries. These include political administration, monitoring and supervision of development projects at the local level and maintenance of community roads. However, not all local governments have the financial or human resources required to provide an adequate level of services. At the lowest levels, much of the work is done by ward councilors. Due to lack of skills, ward councilors poorly facilitate developmental activities.

Efficiency and accountability: lack of transparency and, indeed, corruption are among the spurs for faster devolution. However, part of the reason corruption and abuse of office are so widespread at the center are poor managerial capacities. Devolved corruption should be avoided. This would be especially true where local administration capacities are weaker than at the center. The central government can curb corruption through the use of intergovernmental transfers governed by objective criteria. Conditional grants are disbursed when the local governments fulfil the agreements with the central government regarding certain targets, such as recruiting teachers, purchase of drugs or provision of extension services. These grants may cover recurrent expenditure, whereas unconditional grants cover local capital expenditures. Second, funds for local government operations should not be routed through line ministries. Instead, the Treasury has to remit funds direct to the accounting officer of each provincial and metropolitan province. An innovation would be that the amounts disbursed are published in local newspapers. This would enable the local populations to know what to expect in terms of service provision and acts as a deterrent to financial abuse.

Peace promotion and conflict resolution: devolution addresses territorial grievances. The grievances range from differentiated patterns of income and unemployment to poor infrastructure and low business or economic activity. A major consequential benefit of devolution, therefore, is an “economic dividend” that accrues to regions or territories that are perceived to be disadvantaged by centralized models of development. Devolution is, therefore, a mechanism that facilitates the attainment of territorial justice (Morgan, 2006; Moyo, 2014).

In Matabeleland, there are examples of the Hwange Thermal Power Station and the Kariba Hydro-Electricity Power Station. The majority of the schools and homes in the surrounding Hwange, Kariba and Binga districts have no electricity, yet these two major electricity power stations generate and supply a lot of electricity to the national grid. Furthermore, “three of the country’s busiest border posts (Beitbridge, Plumtree and Kazungula) that remit high amounts of revenue to the state” as well as “several tourist resort centers such as the Victoria Falls, Hwange National Park and Matopos, which similarly remit the highest amounts of foreign currency to the state” (Mhlanga, 2012) are located in Matabeleland provinces yet this region remains one of the poorest and least developed. Given these imbalances and inequitable distribution of revenue derived from exploitation of national natural resources, it is not surprising that calls for devolution were more prominent in these marginalized provinces which anticipated territorial justice through devolution.

Planning as an Instrument of Social Mobilization

The distinctive feature of decentralization and devolution in Zimbabwe is the central role allotted to the planning function of the Provincial and Metropolitan Councils. A comprehensive Provincial and Metropolitan plan is to be prepared by each body before it can claim the grant-in-aid. The planning process starts at the grassroots with maximum involvement of the masses. Mass participation is not limited to elected representatives or voluntary agencies, but includes ordinary people assembling in districts with non-official experts and volunteers participating in preparing reports, formulating projects and drafting plans. Officials are to work alongside non-officials. In order to ensure transparency and participation without compromising technical objectivity, a sequence of phases each with its distinctive objectives, main activities and training program should be drawn up. Mere listing of needs does not make a plan, the resources potential for satisfying them has to be assessed. The second stage may start with a number of participatory studies of the problems and assessment of the local resources, both human and material. The findings of such studies and the deliberations in the districts should be summed up in a development report. Planning is a technical process that involves assessment of needs and resources, fixing of priorities, preparation of projects and formulation of plans. In Zimbabwe, decentralization and devolution planning process should be used as more than technical exercises. They should be used as instruments for social mobilization in support of devolution. The mass mobilization must be intended to ensure that the correct decisions are made at the local level, to bring about certain attitudinal changes among participants and to build up mass pressure from below for democratic reforms.

Devolution Fund Allocation Guidelines

The Constitution recognizes the need for equitable revenue sharing among the tiers of government. It provides for the transfer of fiscal resources from the national government to Provincial and Metropolitan Councils (PMCs) and local authorities. Section 301(3) of the Constitution of Zimbabwe states that not less than 5% of the national revenues raised in any financial year must be allocated to the province and

local authorities as their share in that year. This requires the Ministry of Finance and Economic Development to consider annual revenue projections and determine the percentage allocations to PMCs based on such factors as the population size, poverty levels, infrastructure deficits, socio-economic disparities, national development goals and priorities among others. In addition, Section 301(1) of the Constitution directs Parliament to enact legislation providing for conditional and unconditional grants to PMCs and local authorities. The other imperative for devolution is to share the national cake equally amongst Zimbabwean citizens. The Constitution sets the criteria that inform the allocation of conditional and unconditional grants to provincial and metropolitan councils and local authorities. The devolution policy being crafted by government must spell out the formulae for the allocation of funding to the provinces so that all the regions equally benefit from national resources. However, a comprehensive revenue sharing model should not only be influenced by political parameters. The financial and economic considerations have to be made in determining the weight of population, infrastructure gaps and/or economic disparities in setting the fiscal formulae for revenue sharing in which case, the different provincial and local councils will be funded commensurate to their fiscal needs, effort and gaps.

The Public Finance Management Act needs to be amended to facilitate the allocation of sufficient resources for basic social services and economic development at the local level. The need to implement devolution has subsequently been echoed in various government policy documents. For instance, the Transitional Stabilization Program (TSP) (Government of Zimbabwe, 2018). The TSP provides various measures which the government seeks to implement to change the socio-economic environment in order for Zimbabwe to attain the status of a “prosperous and empowered upper-middle income society by 2030” (Government of Zimbabwe, 2018). A key feature of this plan is devolution of governmental powers, responsibilities and resources with a key focus on provinces. A section of the plan titled *Empowerment of Provinces* reads:

While Zimbabwe remains a unitary State, the implementation of the country’s development programs will allow for devolution to achieve fair and balanced development, spearheaded by Provincial Councils which will initiate development programs for their respective Provinces, consistent with Section 264 of the Constitution (Government of Zimbabwe, 2018).

This represents a new governance framework where decentralization becomes a key feature and strategy for fair and just governance across its four dimensions, namely administrative, political, fiscal and market.

These reforms focusing on the provincial level and devolution at large were also recognized in the 2019 Pre-Budget Strategy Paper published by the Ministry of Finance and Economic Development. The Paper emphasizes the need for local authorities to begin to be more visible in the delivery of services and investment promotion. The fact that the blueprint and the Pre-Budget Strategy Paper were relatively talking to each other, is indicative of the fact that there is the political and administrative will to implement devolution. The 2019 national budget speaks to devolution and the operationalization of provincial councils. The national government allocated US\$310 million to provincial councils in the 2019 financial year as part of the five percent Provincial Councils are entitled to in each financial year. It is not clear whether the omission of metropolitan councils was deliberate, or, whether they were, indeed, factored into this allocation. The actual allocations to provinces require Cabinet approval to ensure that these allocations “target addressing pockets of marginalization in Provinces and Districts” (Government of Zimbabwe, 2018). This requirement is without doubt an indication that while provinces can determine their own development priorities, it is the national government that has the final say since Cabinet can

veto expenditure prioritization proposed by a province. This does not augur well for the principle of devolution, which entails diffusing policy making and implementation powers, including over expenditure decisions, to the local level. These issues need to be clarified in the decentralization and devolution law and policy reforms.

CONSTRAINTS ON LOCAL GOVERNANCE

A fully devolved local government presents an opportunity to solve a number of problems that have been noted within the service delivery system of the country. Some of the challenges noted include too much central governmental interference in local authorities, lack of capacity issues and skills, limited financial resources, lack of accountability and poor management of public services both in the central and local government. However, there are challenges that can undermine Zimbabwe's capacity to implement devolution as enshrined in the Constitution. These are discussed below.

Lack of Citizen Consultations

Local governance can make meaningful impact on people's lives if local communities are given space to participate and provide views on the type of local governance they want. The government is still to give Zimbabweans an opportunity to shape the devolution narrative through devolution discussion.

Unitary State

Historically, Zimbabwe is known to be a unitary state and important decisions on local development have been taken care of by few people mainly from the central government. This poses a serious challenge on the political will to devolve the local development.

Lack of Resources Leads to Program Failure and Reduce Independence

Lack of own resources is probably the single most important impediment to local governance in Zimbabwe. It may limit severely the extent to which local leaders can deliver on their electoral promises, notably improvement of social services, and limits their ability to resist demands and interference from politicians at the center as well as those from line ministries. With their independence, thus, curtailed and lacking their own funds, local governments run the risk of becoming instruments for the propagation of political schemes or for the exercise of patronage by politicians at the center. Lack of resources also limits the extent to which local leaders can motivate their staff and retain them in the now stiff competition between the public and private sectors for educated and efficient employees. A shortage of skilled staff may put a limit on the amount of intervention that local governments can embark on successfully. It also implies that important aspects of accountability, for example, keeping the books in order, refining policies or designing new ones, cannot be undertaken easily. Given that Zimbabwe's educated labor force has expanded substantially since independence, the problem of skills shortage might be resolved when local government design more attractive terms of service, including improved remuneration.

Another concern is on the dwindling revenue base of local authorities with state institutions such as Zimbabwe National Authority (ZINWA), Environmental Management Agency (EMA) and Zimbabwe National Road Agency (ZINARA), taking much of the financial resources that should be going to local

governments in revenue. This raises questions on the sustainability of the financing mechanism to back the process of devolution.

The lack of resources causes a serious shortage of supplies, equipment and physical structures, making daily operations difficult and the realization of policy targets all but impossible. Although lack of basic facilities and tools is partly a result of serious under-capitalization in past decades, the social infrastructure including schools and clinics, destroyed, making their situation more precarious today. Children are still taught under trees in some areas and local government offices are often derelict structures.

Slow Pace in Legislation Alignment

Despite the constitutional provisions that oblige the government to devolve the local governance structures, there is lack of political will from the government's side to enact Acts and legislations that confirm to the constitution's view of devolution. The Urban Council Act [Chapter 29:15] gives excessive powers to the Minister of Local Government and there is too much regulation and interference of the Minister in the operations of the local authorities. The current laws namely the Urban Councils Act and the Rural District Council Act are not aligned to embrace devolution as spelt out in the Constitution.

Center-Local Relations Remain Complex

The central government's wish to control the periphery made nonsense of many earlier attempts to cede powers to local authorities. Direct coercion, including legislation was often necessary to restrain the clamor for autonomy at the local authority level. To what extent, then, the devolution strategy, based on PMCs, would reduce the tension between the center and the councils in matters of policy coordination, budgeting and accountability? The devolution policy should not be hurried, with little preparation of stakeholders, notably the local populations. This may leave considerable room for discretion to politicians. Policy opacity at the local level may make it difficult for the local population to internalize the devolution system. Since the new systems' credibility is depended to a large extent on local people's reaction to its provisions and their ability to penalise poor performance, it is imperative that they understand the issues involved. Devolution threatens various interest groups including politicians. Indeed, whenever there are changes in the political and administrative system there is bound to be resistance and sometimes apathy. These reactions do not mean that change promises little benefit to the population. Since it introduces new ideas and policies and alters the way things are done, change might be the key contribution of devolution development.

But perhaps the most important source of conflict between center and local councils relates to money. Needs-based formula for sharing the national cake needs to be considered. This must take into account the capacity of regions to collect taxes, a degree of inequality in resource distribution remained. The equalization grant is to be based on assessment of resource endowments, with poorer provinces compensated accordingly. However, for the system of transfers to succeed in achieving the goals envisaged by the center and the local government, there is need for a high degree of accountability. This can be achieved through production of monthly accounts of expenditure for purposes of planning and budgeting to the center (Ministry of Finance). Use of threats should be encouraged. The Ministry of Finance must try to reconcile the budgetary process at the center by resorting to threats including withdrawal of funding from provinces not complying with its directives and publishing a list of such provinces in the local newspapers. With devolution, departments in line ministries are also supposed to reduce their functions to those of policy making and supervision, relegating operations and related expenditures to the provincial councils.

Violation of Legislation

The President appointed Provincial Ministers, a move that is in contradiction with the Constitution. There is no constitutional provision that sanctions the appointment of provincial ministers and it is not clear what their duties and responsibilities are. A key question is raised around the operationalization of the provincial councils considering there is going to be an executive minister for the province. It is not yet clear what role these ministers will be playing in light of the provincial councils that the constitution talks to.

SOLUTIONS AND RECOMMENDATIONS

Assistance and Support

Training and capacity building are crucial for successful implementation of decentralization and devolution. Like any other process that seeks to change the status quo, devolution, therefore, requires education. Devolution education is necessary for officials both at the national, PMCs, civil society actors and the general citizenry in the areas of legislation, policies and institutional framework. There is need to enhance capacities of both local community leaders, councilors and their council employees. For example, there is need to carry out capacity building exercises on the duties and functions of councilors. There is also need to strengthen the supply side of local governance through council human resources audits, skills retooling so that it is able to adequately perform its duties.

Public Finance Management System

Sound public finance management systems are a pre-requisite for devolution to be a success and to be meaningful. If the implementation of devolution is hurried without sound public finance management systems, more problems may be created than it seeks to address. When designing a multi-level system of government, it is vital to ensure that there are specific rules, mechanisms and institutions to provide the accountability of various governments to the citizens. Such mechanisms may include independent oversight committees, reporting systems, timeous and easy access to information. Furthermore, the mechanisms include approval of resolutions of PMCs, monitoring visits, performance checklists; for example, service delivery, budgets, compliance, corporate governance and capacity to generate and utilize revenue (Zinyama, 2019).

There is need to have a comprehensive financial monitoring and reporting system. Section 298(1) of the Constitution of Zimbabwe requires clear fiscal reporting at all levels of government. In terms of Section 35(7) of the Public Finance Management Act, within 120 days of the end of the financial year, every provincial council, metropolitan and local authority must submit to the Minister: a) an annual report; b) audited financial statements; and c) the audit report. Training, therefore, becomes a critical component in upgrading competencies and skills of both staff and members.

There is need for political will to address weak internal controls, poor corporate governance, procurement irregularities, and inadequate controls in receipting payments and meeting statutory obligations and deadlines for financial reporting which are recurrent in the successive Auditor General's reports. Without political will to implement punitive measures on errant local and central government officials, the inherent benefits associated with devolution will remain conjectural and theoretical. The first step, therefore, is to enforce the implementation of Section 299(1) of the Constitution regarding the role of

Parliament to monitor and oversee provincial and metropolitan councils as well as local authorities through the alignment of the Public Finance Management Act to the constitution and the promulgation of an Act of Parliament to operationalize devolution. The proposed Act of Parliament should be clear on the obligation of local governments to submit annual budgets and financial performance reports to government, Auditor General and Parliament of Zimbabwe. Such provisions should also incorporate provisions that establish institutional and administrative frameworks that are used by Parliament in monitoring and overseeing state revenues and expenditure.

Politicization of Local Governance in Zimbabwe

There are high levels of infiltration of politics in local governance systems to such an extent that it is difficult for duty holders who are political leaders to dialogue with residents on development issues in a progressive and accountable manner. Failure to address this challenge might affect Zimbabwe's capacity to deliver genuine devolution.

Civic Culture Creation is Fundamental

The PMCs must actively seek to nurture a civic culture that promotes grassroots democratic institutions. A radical transformation of the development culture of the state is a necessary pre-requisite for successful participatory decentralization. It also requires basic attitudinal changes towards the development process among all the key players involved: the elected representatives, officials, experts and the public at large. The bureaucratic departmental approach to development has to give way to an integrated and democratic vision. The ivory tower attitude and deeply ingrained cynicism prevalent among the technical elite will have to be replaced by a culture of participation and engagement. There has to be a transformation in the mindset of the elected representative themselves. Mindset of centralism often stays rooted in the attitudes and behavior of the PMCs beneficiaries, with local authorities appearing tentative and apprehensive in exercising their governance powers.

Political and Administrative Will

Devolution shakes people's comfort zones and the traditional lines of distributing power and resources. Devolution is often a difficult and contested process throughout the world. However, with sufficient political and administrative will, devolution can be implemented.

Long-term Strategy for Devolution Implementation

Evidence throughout the world shows that the implementation of devolution is not an event but an on-going process. The implementation process takes time, requiring review and adjustments from time to time. Importantly, the implementation of devolution requires the backing of the ordinary citizens to ensure sustainability. There is need for think-tanks and academics to continuously inform the process. There is need for a strong center to deliver funding, technical expertise and co-ordination.

Multi-stakeholder Participation

Public participation is *sine quo non* for successful decentralization and devolution. All interested stakeholders including relevant government departments, civil society, residents' associations, PMCs, elected political leaders, church leaders, women organizations, youth organizations, private sector organizations and citizens at large must participate. There is need to develop performance management tools for community leaders, councilors and ministers. The following public participation infrastructure needs to be in place:

Public participation policy: There is need for a consultatively developed policy and framework to guide the public participation process at the Provincial and District levels. This helps to move away from tokenism to real, genuine and effective participation.

Citizen capacity and awareness: both Provincial and Metropolitan Councils must create awareness and capacity build citizens on their rights; and the available opportunities for participation in public financial management. This enhances the quality of inputs from the public in key policy documents such as provincial plans and budgets.

Timely disclosure of information: for fiscal transparency and public participation to be realized, regular and timely disclosure of fiscal data such as strategic plans, budgets, fiscal forecasts and budget monitoring reports among other relevant documents is important. These should be prepared in a simple and clear language for the public.

Timeliness and agenda: The purpose for public engagement should be publicized in advance. The Provincial and Metropolitan entities must clearly specify the scope and agenda of the consultation. The public must be given enough time to internalise the subject matter and develop their submissions. Best practice requires that the notice is given between fourteen (14) and twenty-one (21) days.

Feedback mechanism: The Public should be provided with feedback on their inputs. This requires Provincial and Metropolitan Councils to publish reports of the inputs received from its public consultations and explain how these inputs have been used or otherwise in budget decisions, execution and oversight. Civil society organizations should develop monitoring and evaluation tools for feedback and accountability measurement.

Budget allocation: Provincial and Metropolitan Councils need to budget for public participation. Such allocation assists in awareness creation for meaningful participation.

Thus, the government should support wide consultations in the type of devolution that can work given Zimbabwe's operating environment. There is need for national multi-stakeholder conferences to sensitize and consult citizens on devolution.

Intergovernmental Relations

The structures and processes that governments develop for intergovernmental consultation, co-ordination, and co-operation are referred to as intergovernmental relations. Since Zimbabwe has adopted multi-tiered governance systems, it is necessary that the levels of government communicate with each other. No matter how carefully the powers and responsibilities of each level of government are specified, it is inevitable that there will be areas of overlap in virtually all policy fields. The actions of one government affect other governments. As governments attempt to pursue rational, coherent sets of policy priorities, they realize there is a need to co-ordinate their contacts with other governments. There is need to better understand the implications of the policy decisions, the programs, and the activities of other governments.

A collaborative approach in co-ordination and conflict resolution between or among the three tiers of government should be developed. The relationship must not be confrontational and frigid. Regular meetings are necessary for timely resolution of any differences between or among the levels of government.

FUTURE RESEARCH DIRECTIONS

Inter-governmental fiscal transfers need well established formulas and formulas should be transparent based on credible factors. In this regard, there is need for further research. How do local governments below the provincial level deal with the challenges restructuring of responsibilities and powers between the different levels? What implications does this have for understanding decentralization in Zimbabwe? Other areas include nexus between devolution and intergovernmental relations; public finance management and devolution; and local participation in devolved areas.

CONCLUSION

This paper sought to... It has achieved.... It is concluded that,,

Devolution is not an event but a process that has distinct stages. Decentralization and devolution require a number of changes in administrative structure, allocation of functions and powers and control of resources. All three are interrelated and, to an extent, have to be introduced simultaneously. But it is almost a dictum of devolution that the changes should be effected gradually. Typically, it would be argued that the pre-conditions for successful decentralization and devolution have to be put in place sequentially and with a clear demarcation of functions among the various levels of government. Administrative support structures have to be created by effecting institutional changes, redeploying staff, generating linkages among various agencies and departments.

REFERENCES

- Bossuyt, J., & Gould, J. (2000). *Decentralization and Poverty Reduction: Elaborating the Linkages. Policy Management Brief No.12*. ECDPM.
- Chigwata, T. C. (2018). *Provincial and Local Government Reform in Zimbabwe: An analysis of the Law, Policy and Practice*. Juta.
- Chigwata, T. C. (2019). *Devolution Demystified: Emerging Debates and Prospects for Devolution in Zimbabwe: A Discussion Paper*. <https://zimlii.org/content/devolution/demystified-emerging-debates-and-prospects-for>
- Government of Zimbabwe. (2018). *Transitional Stabilisation Programme and Reforms Agenda: October 2018-December 2020: Towards a Prosperous and Empowered Upper Middle Income Society by 2030*. Harare: Printflow (Pvt) Ltd.
- Government of Zimbabwe. (2018). *Pre-budget Strategy Paper, Towards a Prosperous and Empowered Upper Middle Income Society by 2030*. Harare: Printflow (Pvt) Ltd.

Government of Zimbabwe. (2018). *Address by the President of the Republic of Zimbabwe His Excellency, E.D. Mnangagwa, on the occasion of the State of the Nation Address and the official opening of the First Session of the Ninth Parliament of Zimbabwe*. Harare: Parliament of Zimbabwe.

Government of Zimbabwe. (2018). *The 2019 Budget Speech by Hon. Prof. Mthuli Ncube, Minister of Finance and Economic Development: Austerity for Prosperity*. Harare: 22 November.

Masunungure, E. V., & Ndoma, S. (2013). *The Popular Quest for Devolution in Zimbabwe. Afrobarometer Briefing Paper No. 114, March*. Mass Public Opinion Institute.

Mhlanga, B. (2012). Devolution – the ‘Ticklish’ Subject: The ‘Northern Problem’ and the National Question in Zimbabwe”). *Ubuntu: Journal of Conflict Transformation*, 1(1), 206–215.

Morgan, K. (2006). Devolution and Development: Territorial Justice and the North-South Divide. *Publius*, 36(1), 189–203. doi:10.1093/publius/pjj003

Moyo, P. (2013). *The Devolution of Power Debate and the Zimbabwe National Project. In Nationalism and National Projects in Southern Africa: New Critical Reflections*. Africa Institute of South Africa.

Moyo, P., & Ncube, C. (2014). Devolution of Power in Zimbabwe’s New Constitutional Order: Opportunities and Potential Constraints. *Law, Democracy and Development*, 281-299.

Newsday. (2012). *President Stung on Devolution*. <http://www.newsday.co.zw/14/2/2012-president-stung-on-devolution>

Poverty Reduction Forum Trust. (2019). *Interrogating the New Government’s Vision 2030 and Generating Citizens’ Demands for a shared Economic Reform Agenda. Position Paper*. Poverty Reduction Forum Trust.

The Sundaymail. (2018). *Government Finalises Details for Devolution*. <http://www.sundamail.co.zw/government-finalises-details-for-devolution>

Zinyama, T. (2019). Designing Provincial, Metropolitan Councils. *The Sundaymail*. <https://www.sundaymail.co.zw/designing-provincial-metropolitan-councils>

KEY TERMS AND DEFINITIONS

Decentralization: The diffusion of governmental authority and power away from the national center to other institutions at other levels of government or levels of administration

Devolution: The extensive form of decentralization that diffuses substantial governmental powers, authority, responsibilities, and resources to local units.

Local Government: System of government closest to the people.

Section Autonomy: The extent to which local governments have discretion in carrying out their duties and obligations.

Service Delivery: Socio-economic provision of goods and services such as roads, schools, and hospitals.

A Risk Management on Demographic Mobility of Evacuees in Disasters

Kazuhiko Shibuya

 <https://orcid.org/0000-0002-3243-4757>

ROIS, Japan

INTRODUCTION

To date, our global society is endlessly threatened by disasters (UNISDR, 2018). This article devotes to describe both risks and crises on disaster, and especially it deals with the nature of Fukushima case caused by accident of nuclear power plant (aftermath of the disaster at 11th Mar.2011). Since the human history, this case was one of extreme disasters and multiplied tragedies damaged by quakes, tsunamis, and nuclear power plant accidents. Still now, it presents a number of challenges indeed to be overcome. One of the hardest problems from the inventory of the Fukushima case was characterized as intensive purifications of nuclear pollutions around Fukushima and collective migration of evacuees caused by both natural and human-made disasters. And ongoing disputes against safety seem to vary widely such as compensations, socioeconomic reasons, environmental restorations and community reconstructions. The author intends to review these issues and discuss future design for the lessons from the disaster.

BACKGROUND

No blindness against the truth shall conceal the nature of the serious tragedies around the Fukushima. It should conquer an innermost meaning of sentence “*Ignoramus et ignorabimus*” (We do not know and will not know). Those who independently inquired into the matters ever since the crisis condemned collective optimism and anti-scientism among governmental authorities and polytechnics (Funabashi, Y&Kitazawa, K, 2012). They were hoping to stop more catastrophes, but they clung to saving their faces and dogmatic standpoints. Their statements against risks of nuclear power discredited any citizens. Consequently their poor management of risk prevention in nuclear energy following the tsunamis and nuclear power plant accidents magnified the damage. Their less attention has given to scientific verification ways (Cyranski, D, 2011). Through entirely investigations on consequences after the disaster, Japanese obtained the lessons: *no social systems exist without any scientific verification, and no solutions will be properly conducted without the efforts by true professionalism (academician, governors and other practitioners) involving civic engagements*.

In this point, the SCJ's report pointed out crucially lack of effective risk management for nuclear power plant accidents before the crisis (Science Council of Japan, 2011). Regarding this, in sociology of science, Merton, R.K (1968) already proposed his conceptual term ‘*CUDOS*’, namely *Communitas, Universality, Disinterestedness* and *Organized Skepticism*. Not to mention, scientific community in Japan should be recalled the significance of both Disinterestedness and Organized Skepticism in his prospective discussions. The rule of conduct among scientists as well as anti-crisis operations and imperatives by governors need to be critically reviewed whether national crisis or not. Eventually scientists must be

DOI: 10.4018/978-1-7998-3479-3.ch110

always devoting to strictly scientific governance in their community. Obviously, the social roles shall be undertaken by professional academicians and researchers in terms of their majors standing on the CU-DOS. It is too natural that they shall be obligated to solve above emerged and serious matters. Similarly JST (2011) proposed to solve and tackle those scientific matters by both social and natural scientists. For example, they enumerated modeling and simulation for disaster prevention, risk management, and risk communication for citizens as well as R&D on the nuclear power and new energy technologies (Sugiyama, et al, 2016).

FOCUS OF THE ARTICLE

Governance and Management at Disaster

When governors face at any extreme disaster (including anthropogenic, natural and combining one), they should eagerly devote to achievement of the future restorations (Curt,C.&Tacnet,J.2018). In this point, effective management including evidence based analysis and policy making should be quite needed for actual plans (Biello,D, 2012). Social media analysis also utilizes to unveil communication patterns during disaster (Li, et al, 2014). At least, these concurrent processes are required during disaster as follows.

- Ø *Demographic Data Management:* Migration and Evacuations, Total population of survivors, injuries, fatalities, and etc.
- Ø *Budget Management:* Financing for reconstruction assistances, and etc.
- Ø *Provisioning Management:* Relief supplies, Relief units, Allocating Temporary Evacuation Places, Necessary Resources, Medical cares, Information, Donations, and etc.
- Ø *Access Control Management:* Managing Traffic and Transportations, Restricting area against risks and hazards, and etc.
- Ø *Work Flow Management:* Coordinating total process control, Prioritization on evacuation plans and needs, Project management, and coordinating engagements among stakeholders and volunteers, and etc.

Regarding these points, governors should immediately, responsibly and comprehensively tackle necessary prioritizations. First, it should be tracking dynamics of evacuees and citizens' conditions. Secondly, it offers multiple transportations for evacuees and improves their mobility. Thirdly, it should prepare and allocate temporal housing for them. Of course, without saying, foods and water provision are quite necessary. However, according to revealed data after the disaster, many of municipal governments as well as the national government does not prepare efficiently evacuation plans for large migrants and evacuees. When magnitude 9 level disasters occur around the Tokyo capital area, some researches forecast serious shortages of necessities for two millions migrants and evacuees at least. Those conditions will further cause crisis such as transport congestions, collapse and others. Limitations of budgets and human resources will decisively call exhaustion and confusion in each city. Thus it should make a clearer orientation and plan against future disaster in each municipal city in advance (Batler, D, 2012).

After that, it should turn eyes to persevering purification from the radiation damage caused by the nuclear accidents, its environmental restoration, health monitoring and socioeconomic reconstruction in community (NRC, 2011; IAEA, 2011). As consequently, the Fukushima case definitely needs to solve

nuclear accidents and future design for long-term reconstruction on their devastated communities and hometown. At least, there are including following points.

- Ø *Decommissioning work for the wrecked nuclear power plants in Fukushima*
- Ø *Removal of nuclear fuel and substances*
- Ø *Chemical management on nuclear substances*
- Ø *Purifications in polluted places*
- Ø *Temporary storage and final disposal of contaminated soil, water, and garbage.*
- Ø *Medical health survey and care for victims*
- Ø *Risk assessment against environment and people exposed by nuclear substances*
- Ø *City and Socio-economic reconstructions*

Regarding this concern, enacted provisions often reflect social actualities. For conquests against those hardships of citizens and evacuees, the ‘*Basic Act on Reconstruction in response to the Great East Japan Earthquake (24th, June 2011)*’ enacted the basic policies for reconstructions. For example, a part of provisions clearly stipulated as following.

Article 2. The reconstruction in response to Great East Japan Earthquake will be implemented based on the following

- 1) The unprecedented disaster resulted in enormous damage, where countless lives were lost, numerous people were deprived of their basic living infrastructures and have been forced to evacuate in and out of the disaster-affected regions. Also, the disaster’s influence extends over the entire nation; the economic stagnation in the disaster-afflicted areas is affecting business activities and peoples’ lives nationwide...(hereafter omitted)

At the FY 2014, the total budget of Fukushima Prefecture for reconstruction after the disaster was approximately 1,714.5 billion yen (including both quakes and nuclear disaster countermeasure portion of 870.5 billion yen). And it also includes population declining and aging countermeasures as well as restoration of birth number (142.3 billion yen). In addition, there were 245.5 billion yen for environmental restorations, 144.3 billion yen for living reconstruction assistance, 36 billion yen for expenses to protect medical health of the citizens, 31.6 billion yen for expenses for children and youths who will be responsible for the future, and other expenses.

Further, in addition to above costs, the Fukushima case requires unprovoked compensations and its litigation disputes are ongoing matters in courts (Shibuya,K, 2017). Namely, there remains to solve future designing for reconstruction from devastation in their hometown.

After the disaster, in all things it is better to hope than to despair. But till now, it presents that a number of challenges indeed to be overcome. Nearly future, a decade after the disaster will pass. However both reminiscent verifications and proactive review are required yet (OECD, 2016; Sugiyama, et al, 2016).

With these regards, in this article, especially the author picks up and deliberates focal points as follows. The author formally defines a question. *How is the process of that demographic mobility after the disaster?* and *How to deal with compensation problems?*

1. *Demographic Dynamics after Disaster*
2. *Environmental Restorations and Compensation Principles*

SOLUTIONS AND RECOMMENDATIONS

Demographic Dynamics after Disaster

Principally, the Fukushima case displayed one of the hardest problems related to collective migration of evacuees from Fukushima prefecture caused by both natural and human-made disasters (Akabayashi, A. & Hayashi, Y., 2012). Whereas it is the very significant tasks to track evacuations of citizens at the emergency conditions, but there are often vulnerabilities to be missed. Especially, it should be shed light on forced migration and dynamics of mobility (Butler, D., 2011; Castles, 2003; Jacobse & Landau, 2003; Poston, D. L. & Micklin, M., 2005; Rogers, A., 2015).

At that time (11th Mar. 2011), national authority coercively restricted citizens to stay in wide polluted area around Fukushima nuclear power plants. Then many of ordinary citizens were inevitably destined to escape from Fukushima to other places. At the peak (May 2012), gross migrants from Fukushima (e.g. total population of evacuees) were estimated over 160,000. And including this, gross migrants (e.g. total evacuees in Japan) were estimated over 340,000 at the time of 2012. Of course, both Japan and its municipal governments were eagerly to announce safety information for residents around and nearby Fukushima nuclear plants. But stakeholders were critically accused by many evacuees and citizens, because they did not have any efficient controls and managements. Consequently it entailed severe distrust rooted in information-asymmetry between government and citizens, their risk cognition were imprinting in this way. In addition, this demographic estimation was not too low. Please recall similar cases, for example, the case of Chernobyl in 1986 reported that total population of evacuees was approximately 116,000 around 30 kilometers (INES level 7). And, in 1979, the case of the Three Mile Island accident was estimated over 240,000 around 24 kilometers (INES, level 5).

Ironically, casualties often reminded us a top priority to immediately grasp demographic dynamics after disaster. At that time, the USA government had immediately attempted to understand this tragedy in Japan. NRC (Nuclear Regulatory Commission) (2011) intensively tackled sequential causal analysis by their taskforce organized of nuclear professions. And Global Legal Research Center in the Library of Congress (2013) also already reported above Fukushima disasters and it pointed out the problems on tracking of evacuees and ordinary citizens.

In Japan, Ministry of Internal Affairs and Communications (MIC) and Reconstruction Agency mainly offered demography and disaster relative data¹. But it was unable to obtain data on demographic population and their households' conditions real-timely, and it was just discretely.

In year 2015, a national census every five years was carried in Japan, and it has achieved to unveil many data discrepancies and inconsistencies of population in each local area of Fukushima. Comparing with statistics on resident data holding municipal government office and actual population from census, the latter cases were almost too lower than expected populations in many cases of Fukushima. To date, these missing populations have been not tracing properly and many of evacuees did not intentionally apply immigration cards to municipal government office. It means that many of them still have strong wills to come back to Fukushima in the near future. However the stumbling blocks still remain against their returns, even though the government purifies radioactive substances and pollutions around their towns.

Table 1 shows a part of outflow data on migrants across cities (it includes mobility data within same city). It queried into big-data of MIC (e.g. demographic data of migration and population) and the Geo-spatial Information Authority of Japan (e.g. geospatial data and distance information). The numbers of citizens lived in Fukushima prefecture has been notably decreasing from 2.1 million (2011) to 1.9 million (2015). And this area is statistically 13,780 km² (the third widest area in Japan). And it compares

Table 1. Outflow Data on Mobility Across Cities in Japan (Compared with 2011 and 2015)(Data from Ministry of Internal Affairs and Communications, Japan)

		To								
		Fukushima		Tokyo Area		Nagoya Area		Osaka Area		
From		Sum	Per.	Sum	Per.	Sum	Per.	Sum	Per.	Sum
Fukushima	2011	27,613	0.526	21,112	0.402	1,605	0.031	2,167	0.041	52,497
	2015	1,671	0.686	664	0.273	42	0.017	59	0.024	2,436
Tokyo Area	2011	9,087	0.026	236,446	0.677	37,071	0.106	66,834	0.191	349,438
	2015	581	0.021	22,317	0.800	1,921	0.069	3,077	0.110	27,896
Nagoya Area	2011	588	0.005	40,809	0.355	50,034	0.435	23,535	0.205	114,966
	2015	58	0.008	2,123	0.279	4,272	0.562	1,150	0.151	7,603
Osaka Area	2011	619	0.004	74,869	0.480	22,575	0.145	57,913	0.371	155,976
	2015	83	0.008	3,849	0.368	1,289	0.123	5,235	0.501	10,456

Fukushima with people lived in major metropolis such as Tokyo area (total population is approximately 13 million people and within 9 million people in 23 special districts), Nagoya area (approximately 2.2 million people within central city) and Osaka area (approximately 2.6 million people within central city) in Japan. It namely denotes moving within Fukushima cities, moving towards one of metropolises from Fukushima and moving between metropolises.

Next, according to subsequent progress surveys in Tokyo capital city for evacuees (Shibuya,K, 2017), it found that their motivations which choose the destination and refuge were relying on critical factors: (1) to tie with any kindred relationships (a factor of human relationship), (2) opportunities of jobs in the new address (a factor of new job opportunity), and (3) conveniences to manage their own real estates, farms and factories in their hometown (a factor of holding estates). First factor implies mutual cooperation and helping among local acquaintances, and second answer clearly reflects their needs for job in new dwelling. And third factor relates to geospatial location. Many of citizens adhere to living around Fukushima, and then parts of them have been living in two places of both Fukushima and refuge. Namely, their conditions are back and forth between hometown and refuge. Thus they could not leave to far refuge, and geographical area around 300 kilometers within Fukushima and refuges satisfy their above motivations using transportations such as the bullet train, highways and other land transportations (it drives toward the destination within two hours). For those who required living needs, Tokyo and nearby area could properly offer new job opportunities and dwelling availabilities for them. Regarding this concern, Ravenstein,E.G.(1885) originally discovered ‘*The laws of migration*’, which depended on geospatial distance of collective immigrant workers between each city. Therefore, it is natural to think that this law explains demographic mobility patterns on the Fukushima case.

Environmental Restorations and Compensation Principles

Risk Cognition among Citizens against Pollutions and Contaminations

Nuclear pollution raised the disputes on human rights, health, radiation contaminations to foods and environmental restorations. Both environments (soil, waters and air) as well as people in Fukushima were polluted and exposed by both nuclear substance and radiation (Hayano,R.S et al, 2013; Gibney,E, 2015; Merz,M et al 2015; OECD, 2016). Sampling data accumulated by agricultural scientists endorsed that many of crops absorbed radioactive substances such as ⁹⁰Sr, ¹³⁴Cs, ¹³⁷Cs, and others (Nakanishi,T&Tanoi,K,

2013; Takahashi, T, 2016). And excessive intake and exposure of radioactive pollutants will also endanger citizens' and workers' health (Hiraoka, K, et al, 2015).

Thereafter, the anxiety of citizens against crops and foods was ascribed to above serious radioactive pollutants. Then, they were calling to refrain from purchasing foods and agricultural products made in Fukushima thorough any ways. This social phenomenon often calls severe controversies as a Japanese phrase “*Fuhyo-Higai*”. Of course, these safety matters were incurred temporally. To date, scientific investigations well assured of the safety of agricultural foods. For safety conservation, every crop has been required to pass the food safety examinations, and further the Fukushima prefecture managed the tracking of foods shipments from all Fukushima farmlands to buyers and consumers. But many consumers still uneasily claim skeptical sentiments. Certainly, it didn't subside notable stagnation in local economics. Thus, it is critically true that innocent and honest farmers lost their profits since that time. At least, this social trend was not only engendered by inaccuracy rumor and media communication. Rather, one of main factors was that risk cognition among citizens easily triggers the first domino-effect for disruptions in market and supply chains whether mediated rumor can be clarified or not. These cases indicated that many people did feel something unfavorable and they were not going to buy foods made in Fukushima. And food buyers almost broke the ties of trading and supply networks from suspected as well as trustable farmers, and supermarkets often expelled those agricultural foods (it just labeled and categorized as “Made in Fukushima”).

Externality and its Compensations

With these backgrounds, the focal point of disputes on the compensations has been accused by plaintiff (i.e. citizens, evacuees, and victims) under trials in courts (OECD, 2016). In the Fukushima case, many trials as class action lawsuits are ongoing matters, and these judgments of the first instances are all favorable by the plaintiff. The points of dispute can be summarized as follows. (1) *compensation for health damage to evacuees*, (2) *a situation in which citizens have to evacuate for a long time*, (3) *collapse of the community*, (4) *restoration of the state of origin*, and (5) *execution suspension of the exercise of regulatory by defendants (i.e. National Government and Tokyo Electric Power Co., Inc.)*. Naturally their legal battles may continue, but law courts will almost uphold plaintiff. Further total amount of those compensations can be easily expected to exponentially rise up. And then, with reducing the burden against the gross domestic product in Japan, it is quite important to recover the sales networking, restore the productive capacity of the afflicted area, address the economic damage of producers and make efforts for future economic development. At the time of 16th October 2013, the Board of Audit of Japan estimated the total burden against the compensations in the Fukushima nuclear power plant, and its amount will exceed five trillion yen and more. Additionally, they also simulated that it will take more 31 years to repay the full amount.

Furthermore, it should be also pursued these matters including environmental goods (i.e. non-market goods), its assessment and value restoration. As the first step, there are needs to formal assessments and estimate economic utility of environment itself. Please recall the US NOAA Report (1993), it is indispensable to find the policy problem by evaluation on environment itself and to explore the steps for improvement. Here, before the disaster at 2001, Fukushima Prefecture had already estimated their utility valuations holding environmental real estates (e.g. forests, mountains, rivers, etc.) by travel cost methods as one of revealed preferences methods. According to them, they estimated total value of both ‘flood prevention function’ and ‘air purification function’ was about 286 billion yen. In the breakdown, the value of 55.2 billion yen was calculated for ‘health rest and relief function’. From this point of view,

anyone can clarify the transition of value before and after the disaster (Not to mention, *at least* they lost those values by the disaster).

In the first place, if the damage occurred on the basis of the economic activity of the company, its actual damage called '*negative externalities*' in economics. If the damage was caused by a kind of anthropogenic disasters, the criminal liability is socially relying on the economic activity of the enterprise, which is a result of having a negative effect on the citizens living in the area where the disaster occurred. Certainly, it is often very difficult to coordinate with social welfare functions among stakeholders (Arrow, 1963; Sen, 1970). As well-known as 'Kaldor-Hicks compensation principle' (Hicks, 1939; Kaldor, 1939), it has been discussed in the field of welfare economics, it was damaged due to the externalities by corporate economic activities, and the focal point is on how they compensate for their victims. However, when each preference between the company and the victim can be defined, it has been proofed that their social choice in society does not necessarily reach to the Pareto optimum. As for this problem, Sen (1970) strictly deliberated that the Kaldor-Hicks compensation principle prevails in the process leading to the social choice based on preference order is limited, and the defect was resolved by the new compensation principle by Scitovsky (but Skitovsky's proof had also been shown to have drawbacks, which has been critically revised by Sen). Meanwhile, in Fukushima cases, many of lawsuits said that compensation amounts are unequal conditions among victims. Economically speaking, so far, Sen also asserts that the assessment of inequality in economics has been neglected in economics, and he taken up as an example of social loss due to Atkinson's equivalent income. As taking into account the diversity of individuals, he definitely defines inequality.

Regarding this context, the Coase theorem (Coase, R.H, 1960) defines to fulfill three conditions such as (1) *low negotiation costs*, (2) *clarity of attribution of rights*, and (3) *endorsement for completeness of information*. However, there is no guarantee that nuclear accidents will satisfy these conditions. In particular, the completeness of information is not guaranteed in the Fukushima case. Moreover, regarding negotiation and transaction costs, risk assessment and safety inspection expenses will increase according to the damage. In addition, the impact of secondary damage, the disappearance of regional brands as 'public goods' and the repair costs will not necessarily be ruled in this theorem. As high level expertise in engineering and related fields is required, information asymmetry among stakeholders can not still fill in the gap between them. For these reasons, the actual solution will not be easy as the theorem assumes.

FUTURE RESEARCH DIRECTIONS

All things considered, the Fukushima case impressively implied us the significance of maintaining habitable zones and environmental restorations damaged by nuclear accidents. Development of human's civilization was coeval with an inevitable and strong problem on interlinks between population and geospatial environment (DeMenocal,P.B&Stringer,C, 2016). Loss of control will be a menace to our sustainable society. In the context of the four priorities for actions based on the Sendai Framework for disaster reductions, it is certainly that various methodologies and in-depth ideas were quite needed. Especially, simulation methodology offers the benefits for stakeholders (Meadows, D.H et al, 2004; Shibuya,K, 2004, 2017). It could demonstrate some base scenarios using various environmental, socio-economic and demographic factors. As integrating with systematic concerns, it is possible to simulate the interlinked problems. The further studies should unveil multiple influences between causes and results for resilience of communities and socioeconomic reconstruction.

First, many citizens lost their tolerance of unsafety when governors trampled on their daily lives, and then they left Fukushima. Restoration of the natural environment due to radioactive contamination was a serious problem in itself. As stated by outside stakeholders, pollution on agricultural crops became noticeable. So even if some residents returned home and the community was physically rebuilt, as long as agriculture is the core of the regional economy, their reconstructions will be still difficult. In fact, there are many citizens who settled in evacuation outside the prefecture, and gave up the return.

Secondly, for next several decades, it should assess some scenarios based on the time discount principle in Fukushima. Can stabilize economic valuations between contemporary and future generations? As long-term reconstructions, any present value does not have equal value in the future. There are the necessities to examine detail dynamics and linkages in socioeconomic and demographic factors. Through intensive investigations, it will unveil multiple influence factors against disaster resilience and socioeconomic reconstruction. For instance, when temporal migrants and evacuees finally become long-term residents and settlers in city, each city should reconsider socioeconomic balances for acceptance of migrations. It is similar to the migration policy studies in economics (Powell, B., 2015). Increasing active workers' population raises economic benefits for accepting city, but retired generations require their pensions and social welfare services.

Thus for future generations in Fukushima, the basic design on these matters should not be reshaped by a standpoint of contemporary generations. It remains to solve the problems including a perspective of the future generations. Their consensus should be depended on future plan (Johnson, G.F., 2008). It prioritizes more future common wealth than present value based on individual's preferences. However, it will be perhaps the long way to achieve their goals.

CONCLUSION

The goal of this article was to review the nature of the Fukushima disaster. The author deems the necessary to further contemplate those matters for risk and crisis management. This case presents a number of challenges still to be overcome. And it should be deepened by future studies for envisioning the sound design in regional communities.

REFERENCES

- Akabayashi, A., & Hayashi, Y. (2012). Mandatory evacuation of residents during the Fukushima nuclear disaster: an ethical analysis. *Journal of Public Health*, 34(3), 348–351.
- Arrow, K. J. (1963). *Social Choice and Individual Values*. Yale University Press.
- Biello, D. (2012). *It's not just Fukushima: mass disaster evacuations challenge planners*. <https://www.nature.com/news/it-s-not-just-fukushima-mass-disaster-evacuations-challenge-planners-1.10165>
- Butler, D. (2011). *Reactors, residents and risk*. <https://www.nature.com/news/2011/110421/full/472400a.html>
- Butler, D. (2012). France 'imagines the unimaginable'. *Nature*, 481(7380), 121–122. doi:10.1038/481121a PMID:22237081

Castles, S. (2003). Towards a Sociology of Forced Migration and Social Transformation. *Sociology*, 77(1), 13–34. doi:10.1177/0038038503037001384

Coase, R. H. (1960). The Problem of Social Cost. *The Journal of Law & Economics*, 3, 1–44. doi:10.1086/466560

Curt, C., & Tacnet, J. (2018). Resilience of Critical Infrastructures: Review and Analysis of Current Approaches. *Risk Analysis*, 38(11), 2441–2458. Advance online publication. doi:10.1111/risa.13166 PMID:30102784

Cyranoski, D. (2011). Quake shuts world's largest nuclear plant. *Nature*, 448(7152), 392–393. doi:10.1038/448392a PMID:17653149

DeMenocal, P. B., & Stringer, C. (2016). Human migration: Climate and the peopling of the world. *Nature*, 38(7623), 49–50. doi:10.1038/nature19471 PMID:27654915

Funabashi, Y., & Kitazawa, K. (2012). Fukushima in review: A complex disaster, a disastrous response. *Bulletin of the Atomic Scientists*. <http://bos.sagepub.com/content/early/2012/02/29/0096340212440359>

Gibney, E. (2015). *Fukushima data show rise and fall in food radioactivity: Giant database captures fluctuating radioactivity levels in vegetables, fruit, meat and tea*. <https://www.nature.com/news/fukushima-data-show-rise-and-fall-in-food-radioactivity-1.17016>

Goncalves, B., & Perra, N. (Eds.). (2015). *Social Phenomena: From Data Analysis to Models (Computational Social Sciences)*. Springer. doi:10.1007/978-3-319-14011-7

Hayano, R. S., Tsubokura, M., Miyazaki, M., Satou, H., Sato, K., Masaki, S., & Sakuma, Y. (2013). Internal radiocesium contamination of adults and children in Fukushima 7 to 20 months after the Fukushima NPP accident as measured by extensive whole-body-counter surveys. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, 89(4), 157–163. doi:10.2183/pjab.89.157 PMID:23574806

Hicks, J. (1939). The Foundations of Welfare Economics. *Economic Journal (London)*, 49(196), 696–712. doi:10.2307/2225023

Hiraoka, K., Tateishi, S., & Mori, K. (2015). Review of health issues of workers engaged in operations related to the accident at the Fukushima Daiichi Nuclear Power Plant. *Journal of Occupational Health*, 57(6), 497–512. doi:10.1539/joh.15-0084-RA PMID:26269280

Jacobse, K., & Landau, L. B. (2003). The Dual Imperative in Refugee Research: Some Methodological and Ethical Considerations in Social Science Research on Forced Migration. *Disasters*, 27(3), 185–206. doi:10.1111/1467-7717.00228 PMID:14524045

Johnson, G. F. (2008). *Deliberative Democracy for the Future: the case of nuclear waste management in Canada*. University of Toronto Press. doi:10.3138/9781442687837

Kaldor, N. (1939). Welfare Propositions of Economics and Interpersonal Comparisons of Utility. *Economic Journal (London)*, 49(195), 549–552. doi:10.2307/2224835

Kasperson, J. X. (2003). The Social Amplification of risk. In N. Pidgeon, R. E. Kasperson, & P. Slovic (Eds.), *The Social Amplification of Risk*. Cambridge University Press. doi:10.1017/CBO9780511550461.002

- Li, J., Vishwanath, A., & Rao, H. R. (2014). Retweeting the Fukushima Nuclear Radiation Disaster. *Communications of the ACM*, 57(1), 78–85. doi:10.1145/2500881
- Meadows, D. H. (2004). *The Limits to Growth*. Chelsea Green Pub.
- Merton, R. K. (1968). *Social Theory and Social Structure (1968 Enlarged Edition)*. Free Press.
- Merz, M., Shozugawa, K., & Steinhäuser, G. (2015). Analysis of Japanese Radionuclide Monitoring Data of Food Before and After the Fukushima Nuclear Accident. *Environmental Science & Technology*, 49(5), 2875–2885. doi:10.1021/es5057648 PMID:25621976
- Nakanishi, T., & Tanoi, K. (Eds.). (2013). *Agricultural Implications of the Fukushima Nuclear Accident*. Springer. doi:10.1007/978-4-431-54328-2
- Pidgeon, N., Kasperson, R. E., & Slovic, P. (Eds.). (2003). *The Social Amplification of Risk*. Cambridge University Press. doi:10.1017/CBO9780511550461
- Poston, D. L., & Micklin, M. (Eds.). (2005). *Handbook of Population*. Springer.
- Powell, B. (2015). *The Economics of Immigration*. Oxford University. doi:10.1093/acprof:oso/9780190258788.001.0001
- Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London*, 48(2), 167–235. doi:10.2307/2979181
- Rogers, A. (2015). *Applied Multiregional Demography: Migration and Population Redistribution*. Springer. doi:10.1007/978-3-319-22318-6
- Sen, A. (1970). *Collective Choice and Social Welfare*. Harvard University Press.
- Shibuya, K. (2004). A Framework of Multi-Agent Based Modeling, Simulation and Computational Assistance in an Ubiquitous Environment. *SIMULATION: Transactions of The Society for Modeling and Simulation International*, 80(7), 367–380. doi:10.1177/00375497040406740
- Shibuya, K. (2017). *An Exploring Study on Networked Market Disruption and Resilience*. KAKENHI Final Research Report. (in Japanese)
- Sugiyama, Sakata, I., Shiroyama, H., Yoshikawa, H., & Taniguchi, T. (2016). Research management: Five years on from Fukushima. *Nature*, 531(7592), 29–31. doi:10.1038/531029a PMID:26935680
- Takahashi, T. (Ed.). (2016). *Radiological Issues for Fukushima's Revitalized Future*. Springer. doi:10.1007/978-4-431-55848-4

ADDITIONAL READING

- Library of Congress. (2013). Japan: Legal Responses to the Great East Japan Earthquake of 2011, <http://www.loc.gov/law/help/japan-earthquake/Great-East-Japan-Earthquake.pdf>
- NOAA. (1993). Report of the NOAA Panel on Contingent Valuation, <https://www.researchgate.net/publication/235737401/download>

OECD/NEA. (2016). Five Years after the Fukushima Daiichi Accident: Nuclear Safety Improvements and Lessons Learnt. <https://www.oecd-neo.org/nsd/pubs/2016/7284-five-years-fukushima.pdf>

Science Council of Japan. (2011). Report to the Foreign Academies from Science Council of Japan on the Fukushima Daiichi Nuclear Power Plant Accident. <http://www.scj.go.jp/en/report/houkoku-110502-7.pdf>

Sendai Framework for Disaster Reduction. <https://www.unisdr.org/we/coordinate/sendai-framework>

UNISDR. (2018). Disaster Displacement: How to Reduce Risk, Address Impacts and Strengthen Resilience. https://www.unisdr.org/files/58821_disasterdisplacement05a.pdf

KEY TERMS AND DEFINITIONS

Coase Theorem: Coase theorized economic balance between efficiency and outcome caused by externalities among stakeholders in economics (Coase, 1960), and it defines to fulfill three conditions such as (1) low negotiation costs, (2) clarity of attribution of rights, and (3) endorsement for completeness of information.

Kaldor-Hicks Criterion: This criterion refers to discussions on compensation principle in welfare economics (Hicks, 1939; Kaldor, 1939), and it also widely applied in policy studies and environmental studies. It characterizes the form of cost–benefit analysis. In the cost–benefit analysis, each policy is deliberated and evaluated by comparing the total costs (e.g., building costs and environmental costs) with the total benefits (e.g., profits and convenience for stakeholders).

Migration and Evacuation in Disaster: Aftermath of the serious disasters, it should be shed light on migration and demographic dynamics of evacuees. Certainly, it should distinguish between forced migration and spontaneous one (Castles, 2003; Jacobse & Landau, 2003). Formally, the former can be considered that governors and officers in charge can enact necessary legal imperatives for citizens, and the latter can be based on the will of each citizen. However, human-made disaster usually generates inevitable matters against unneglectable risks, and then every stakeholder directly faces the evacuation problems.

ENDNOTE

¹ <http://www.stat.go.jp/english/index.html>

GIS–Based Quantitative Landslide Risk Assessment Approach for Property and Life at Bartın Hepler

Arzu Erener

Kocaeli University, Turkey

Gülcan Sarp

Süleyman Demirel University, Turkey

Şebnem Düzgün

Colorado School of Mines, USA

INTRODUCTION

Landslides are the second most common natural disasters that cause damages in Turkey after the earthquakes (Ergunay et al., 2003). 89% of the middle and the eastern parts of the Black Sea Region are reported to be susceptible to landslides (Toprak Su 1978; Öztürk 2002). Thus, landslide risk assessment is of crucial importance in decreasing the potential losses.

However, attempts to determine qualitative and/or quantitative risk assessment seems to be fewer (e.g. Wang and Unwin 1992; Fell 1994; Chung et al. 1995; Carrara et al. 1995; Jibson et al. 1998; Aleotti and Chowdhury 1999; Chung and Fabbri 1999; Chowdhury and Flentje 2002; Dai et al. 2002; Bell and Glade 2004; Remondo et al. 2008; Erener and Düzgün 2010; Silva and Pereira, 2014; Luna et al., 2014; Erener, et al., 2016, Erener et al. 2017; Pereira et al., 2017; Vega and Hidalgo, 2017; Hidalgo et al., 2018; Vega et al., 2019) which might be due to the difficulties in obtaining data.

One of the chief advantages of a Quantitative Risk Assessment (QRA) is that it can be compared with other types of risk that can affect a community and since of its quantitative nature it can be communicated more comprehensible to the policy and decision-makers to be used for risk management policies (Luna, et al., 2014). In recent years, Geographical Information Systems (GISs) and Remote Sensing (RS) have proven to be common tools adopted for various studies. In this study, a GIS-aided quantitative landslide risk assessment approach is proposed with an application to local-scale.

BACKGROUND

Quantitative Risk Assessment (QRA) requires quantification of losses due to a given landslide hazard. When landslide hazards are investigated in a local scale, it is necessary to predict the run-out area to identify the potential damages to elements at risk. In any QRA it is of fundamental importance to identify and characterize elements at risk (e.g. buildings, roads, utilities, lifelines) for predicting losses. If insufficient or no digital data exist on a local scale, remote sensing products can be used to extract the element at risks, such as road networks and buildings. Most of the works in the literature concentrates only on the extraction of a single object such as only buildings or only roads. Typical examples of early

DOI: 10.4018/978-1-7998-3479-3.ch111

works in building detection are Lin and Nevatia (1998), Kim and Nevatia (1999), Peng and Jin (2007), Vakalopoulou et al. (2015). On the other hand, Klang (1998), Laptev et al. (2000), Amini et al. (2002), Mena and Malpica (2003) Christophe and Inglada (2007), Yang and Wang (2007), Aytekin et al. (2012) used the most common algorithms for the detection of road.

The population data is required for each individual building to estimate the loss of life due to landslides for risk assessment studies. The population data can be collected through surveys that are time-consuming and costly. On the other hand, recently RS has proved to have the potential for predicting the population with low cost and up-to-date data. Zhang (2003), Lu et al. (2006), Morton and Yuan (2008), and Erener and Düzgün, (2009) has examined residential population dynamics from satellite remote sensing images.

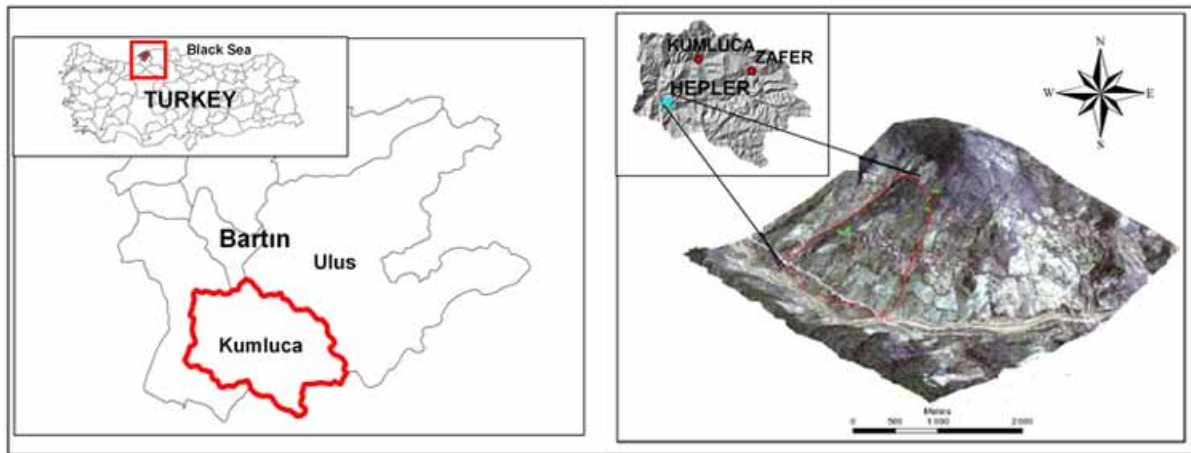
Vulnerability is another fundamental component in QRA (Leone et al., 1996). Reviews of landslide vulnerability are made by Glade (2003) and Lee and Jones (2004). The vulnerability of static elements at risk, such as buildings, land-use, roads, etc. does not vary with time. They are always exposed to threats for all time in all day or all year. On the other hand, the vulnerability of dynamic elements at risk may change depending on their geographical and temporal existence with respect to landslides and it is more difficult to estimate the vulnerability for these elements. Thus, for dynamic elements (residents of the buildings, cars, trains, etc.) daily and seasonal movements are also required for a reliable estimation. Quantification of elements at risk requires a compilation of data from different organizations through interviews with the employees, which is a time-consuming process. Therefore, the assessments of vulnerability for landslides in the literature are usually qualitative (Rautela and Lakhera, 2000) and mostly depend on historical records (Dai et al., 2002; Glade, 2003). For effective use of QRA, generalized quantitative models for vulnerability assessment are essential. Several procedures are described in the literature for the assessment of vulnerability, (e.g. Mejia-Navarro et al. 1994; Fell 1994; Leone et al. 1996; Glade 2003; Uzielli et al. 2006).

FOCUS OF THE ARTICLE

This study introduces a GIS based approach for QRA and illustrates its implementation to Hepler village in Bartın Province, Turkey. GIS is very modern and useful data entry, processing, analysis, query and update tool for both spatial and attributes data (URL1) used at whole parts of this study. On the other hand, RS is an effective data collection method for GIS. A generic algorithm (Aytekin, et al. 2012) that consider road and building extraction together from remote sensing images is adopted in this study for detection of elements at risk. Landslide run-out models being used for assessing the vulnerability values quantitatively on local scale. The velocity of landslide in its run-out area and the number of inhabitants per house are taken into account to quantify the vulnerability related to the elements at risk at GIS environment. In contrast to the property, human exposure to the landslide occurrence may vary depending on the temporal impact. Therefore, the temporal impact is considered for the residents living in buildings. In this context, the occupancy of buildings by people at different times is considered in the study. Hence, the vulnerability of human life is evaluated for temporal scales of day time, night time, high season time, and low season time. Furthermore, the vulnerability of buildings and road are estimated by adopting the damage probability matrix approach by Düzgün (2008). This approach is used in earthquake engineering for evaluating the damage of building stock for a given earthquake intensity (Ko Ko et al., 2004).

As a result of the analysis, quantitative risk maps for both life and property were obtained by the evaluation of hazard, vulnerability and consequence maps for each element at risk at GIS environment. A combined landslide risk map was obtained by adding risk for each element at risk on a continuous

Figure 1. Location map of the study region



scale where numerical values indicate the distribution of risk including the annual probability of expected losses in \$ per pixel and the annual probability of life loss per pixel for property and life, respectively.

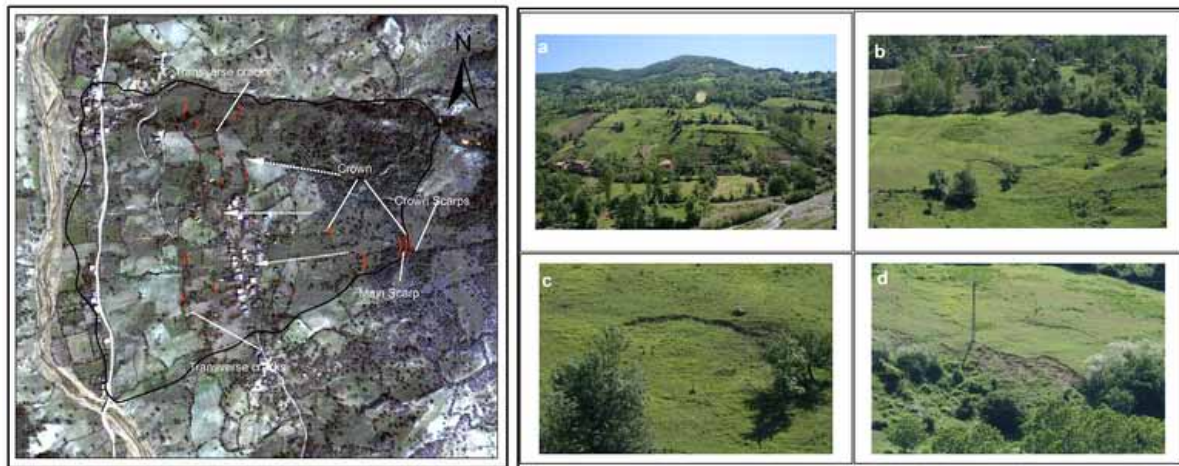
METHOD OF STUDY

Hepler is a small village located on a hillside. It exhibits mountainous topographical features and is frequently subjected to heavy precipitation. Due to these adverse effects, the village is prone to extensive and severe landslides. There are approximately 45 buildings, many of which are constructed on the slope with one and two stories. The population in Hepler village is approximately 250, which may change during summer or winter. The village covers 0.5 km² and is located approximately 8 km southwest of Kumluca town. The 3D view of the study region is given in Figure 1.

The topographical elevation varies from 241 to 527 m above sea level. The predominant slope angle is approximately 24° with slope angles ranging from 0° to 65°. Depending on the reports of the General Directorates of Disaster Affairs of Bartın, the settlement is on an old landslide mass. Landslide mass is identified in lithologic units of Upper Cretaceous age Flysch, which is named as the Uluş formation. The Flysch formation is comprised mostly of thick sandstones in the lower parts, then sandy, clay schist, clay marl, and claystone units on the upper parts. The average soil thickness varies between 30 and 40 cm. The Flysch formation is saturated due to extreme snowmelt in the winter and rainy season in the spring. It is reported that it rains above average during seasonal rains and it snows more than 1.5 m high. The snow melts slowly and as a result, the terrain covered with clay absorbs the whole water. As the terrain becomes saturated with water, landslide and mudflows occur.

In order to determine a detailed topography of the slope and assess the properties of the landslide, extensive fieldwork was carried out. The Ashtech Kinematic GPS (Global Positioning System) was used to collect the X, Y, and Z coordinates of the surface topography with an accuracy of ± 0.5 m (at 95% confidence interval). When terrain conditions were not suitable to collect data points with GPS because of the dense vegetation cover or steep slopes, TruPulse 200 Laser Range Finder was used. The measurement range of the range finder is up to 3280 ft (1000 m) in distance and ± 90 degrees as inclination. It has an accuracy of ± 1 ft (± 30 cm) in distance and ± 0.25 degrees as inclination. The data points were collected homogeneously from the whole slope and as a result, the DEM was constructed at 1 m

Figure 2. The cracks and scarps observed and mapped through field survey at 6-7 June 2009 and the photos of the terrain with scarps and crowns



grid size, using a Triangular Irregular Network (TIN) approach. The observed cracks and scarps through the field survey were mapped onto the remote sensing image of the study region (Fig. 2).

As presented in Fig. 2, in the southern part of the region, transverse cracks were observed. Moreover, the main scarp boundary and crowns were observed at the top of the slope. Some of these cracks are shown in photos given in Fig. 2 a, b, c and d. The crowns observed through the study region were mostly 30-40 cm deep. Depending on the interviews with the local people, these artifacts occurred between 2008 and 2009. Thus, these fractures indicate that the slope mass is still active and slowly moving.

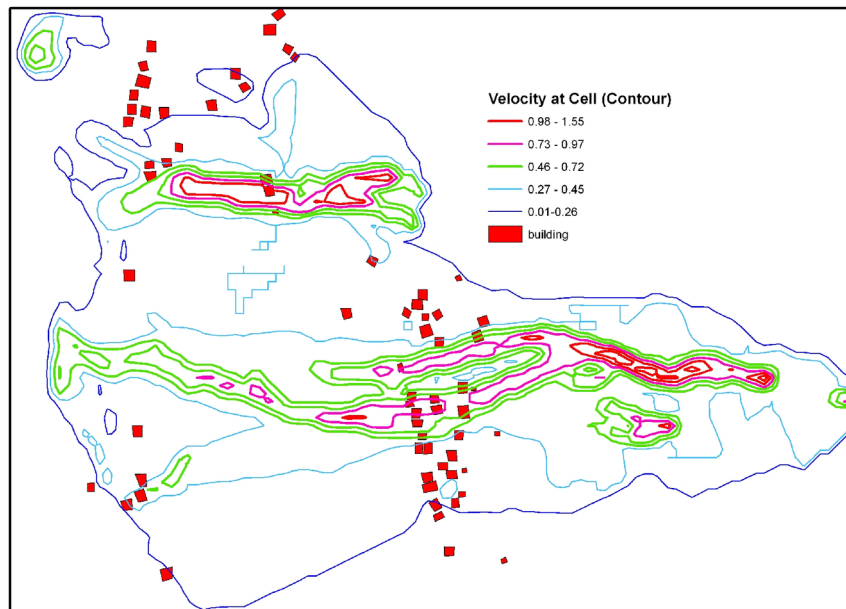
Consequence Analysis

On a local scale, run-out area is critical in identifying the elements at risk and associated vulnerabilities to analyze the consequences of a landslide. A physical process model (Flo 2D) was used for modeling the run-out. Flo 2D is a two-dimensional finite-difference model that simulates clear water, flood hazards, mudflows, and debris flows. The objective of this model is to estimate the probable range of flow properties in terms of velocity and depth to predict a reasonable area of inundation O'Brein et al. (1993) and O'Brein (1999). The details of the modeling the run-out was presented at (Erener, 2009). The model provides the velocity of the debris flow so that the vulnerability of the buildings and also the vulnerability of the people living in the buildings could be predicted depending on their exposure to slide velocity. The velocity map of the simulated slide changes with the volume concentration and ranges between 0.01 and 1.55 m/s. The spatial variation of velocity values over the study region and the distribution of houses are illustrated in Fig. 3.

The distribution of the elements at risk in space and time was extracted from the Quickbird image for the consequence analysis. The details of the image analysis algorithm were presented at Aytakin et al. (2012). The elements at risk in the considered region mainly involve buildings, roads, and populations. Fig. 4 shows the delineated candidate buildings detected for each sub-region with the proposed algorithm and overlaid over the manually labeled ground truth.

Road objects are significant infrastructures for risk analysis. Thus, the roads were also considered in the study for the vulnerability analysis and extracted automatically by the developed algorithm from the

Figure 3. The velocity map of the simulation result (m/s)



high resolution remote sensing image (Fig. 5). The elements at risk obtained from the Matlab environment were then assigned to GIS environment for further risk analysis.

Vulnerability Assessment

The degree of damage in buildings and road, (vulnerability) can be estimated by adopting a damage probability matrix (Yüçemen, 2002; Düzgün, 2008). This approach was used in earthquake engineering for evaluating the damage for a given building stock at a given earthquake intensity (Ko Ko et al., 2004). In this approach, instead of building blocks, the buildings and the road were considered separately by the damage probability matrix. Damage to buildings and the road was categorized into four groups based on the degree of damage. Damage State (DS) was considered in four levels, changing from none to destruction, in which none and destruction stand for no and complete damage. The Range of Damage

Figure 4. The candidate buildings extracted at the end of the algorithm overlaid on the ground truth

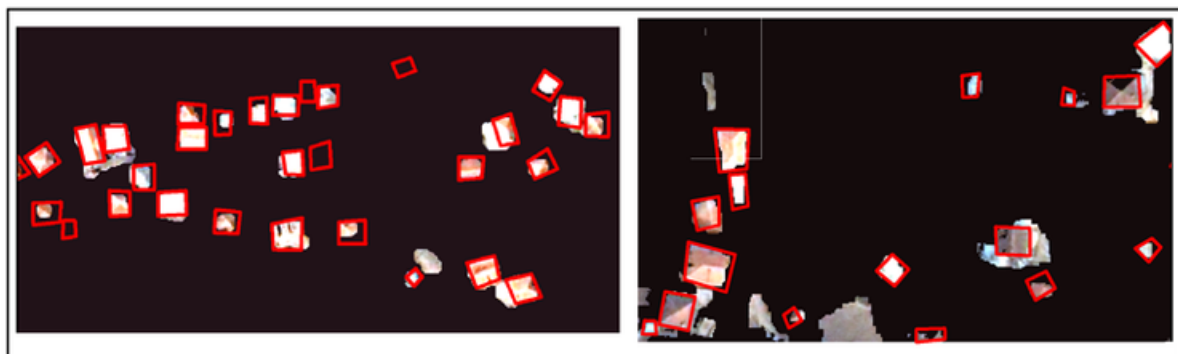
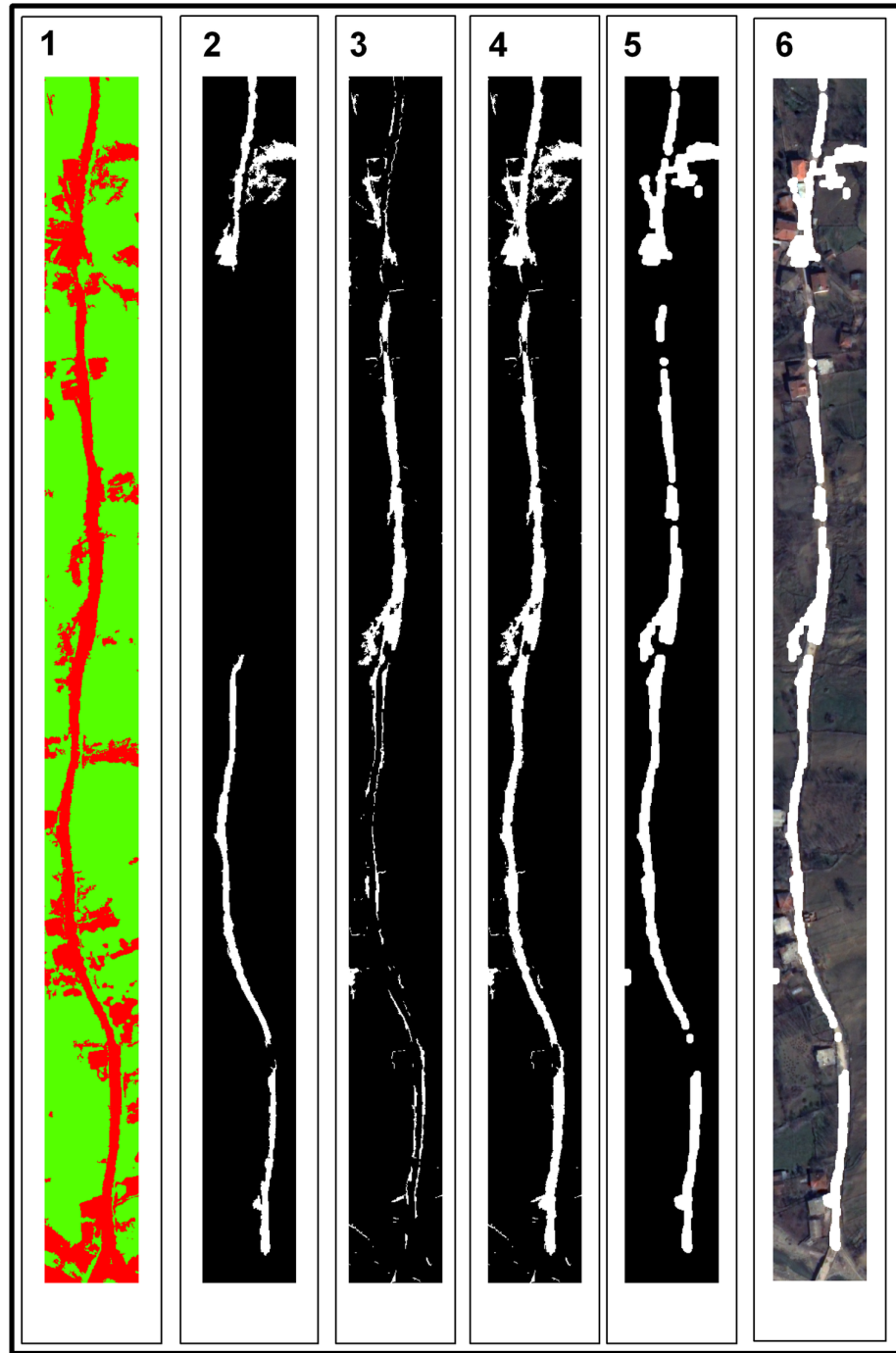


Figure 5. The road detection outputs in the sequence where 1: The vegetation and shadow mask; green is the vegetation and shadow mask, red is the man-made features (road and buildings) 2: The resulting map obtained from the linear feature detection algorithm. 3: The resulting map after PCA algorithm. 4: The resulting map of both linear feature detection and PCA is overlaid to obtain the road feature 5: The morphological processes (imfill open, erode, close) implemented to road feature 6: The road feature overlaid to the remote sensing image



Ratio (RDR) for each damage percentage is defined as 0, 0-10, 20-40 and 50-100 (Düzgün, 2008). Central Damage Ratio (CDR) is defined as the mean value of RDR. The probability of damage state (SPD) for the given temporal condition (season in a year and time in a day) is determined. The properties (such as buildings, roads, etc.) are always exposed to threats all the time. Hence the temporal probability is not considered for the vulnerability analysis of properties. SPD is computed by considering the spatial vulnerability of buildings, which is determined by modeling the landslide velocity using Flo 2D. In the first step the velocity was normalized to 0 and 1. In the second step, the velocity map was classified into four different classes depending on the natural break method. Finally, the percentage area of each property on each level of landslide velocity was computed by the overlay of the velocity map with buildings and roads, respectively. The SPD is the percentage ratio of each feature computed for buildings and roads on each velocity level (Table 1).

Table 1. Damage Probability Matrix for Buildings and Roads

Damage Probability Matrix for Building				Damage Probability Matrix for Roads		
DS	RDR (%)	CDR (%)	SPD	RDR (%)	CDR (%)	SDP
None	0	0	0.2	0	0	0.71
Light	0-10	0.05	0.5	0-10	0.05	0.11
Moderate	20-40	0.3	0.1	20-40	0.3	0.19
Destruction	50-100	0.75	0.1	50-100	0.75	0
Mdr (%)			15			6.08

The MDR is computed for each damage state by sum of the multiplication of CDR and SPD values as presented in Eq. 1

$$MDR = \sum_{j=1}^4 CDR_j \times SPD_j \quad (1)$$

Where j is the level of damage state that ranges between 1 and 4, in which 1=None, 2=Light, 3=Moderate, and 4= Destruction

The damage probability matrix was created by considering the occupancy of people in the building. In contrast to the property, human exposure to the landslide occurrence may vary depending on the temporal impact. The day and night and also the summer and winter should be considered for the assessment of consequences. Hence, depending on the interviews with the local people, the occupancy of each building is found to be between 3 and 10 for low and high seasons, respectively. Seven months of a year (October, November, December, January, February, March, April) were considered as the low season (LS) and the other five months were considered as the high season (HS) (Düzgün, 2008). The probability of landslide occurrences in low (PLS) and high seasons (PHS) was then computed as 0.58 (7/12) and 0.42 (5/12), respectively.

In addition to the time component, the spatial vulnerability of each building should also be considered. Depending on the simulated velocity of landslide, the vulnerability of each building was determined.

Depending on the simulated landslide velocities and the survey data, a quantitative SPD was defined. SPD values are consistent with each other, where MDR for HSNT has the highest value, while MDR

for LSDT has the lowest one (Table 2). Damage state (DS) was considered in three levels which involve none, injury, and fatal. The range of damage ratio (RDR) for each damage percentage was defined as 0, 10-80 and 80-100 by adopting the assumption of Düzgün, (2008).

Table 2. Damage Probability Matrix for Loss of Life

DS	RDR (%)	CDR (%)	HSDT	HSNT	LSDT	LSNT
None	0	0.00	0.44	0.22	0.41	0.34
Injury	10-80	45.00	0.53	0.27	0.49	0.41
Fatality	80-100	90.00	0.10	0.51	0.03	0.25
Mdr (%)			32.87	57.64	24.74	40.95

SOLUTIONS AND RECOMMENDATIONS

Risk Assessment

Risk assessment consists of integrating loss analysis with hazard analysis. Landslide hazard analysis requires determining the spatial and temporal probability of occurrence of damaging events (Bell and Glade, 2004). The hazard maps created for Kumluca basin described in detail by Erener and Düzgün (2012) were used to obtain the hazard values for Hepler village. The area for the Hepler village in the regional hazard map was selected to identify the local risk (Fig.6a).

After the determination of the hazard values from the regional scale hazard map, the risk was estimated for Hepler in the GIS environment. The risk to property was obtained for Hepler by considering the roads and buildings on local scale. The cost of buildings, which is determined by the General Directorate of Disaster Affairs of Bartın, was used to assign the economic value, which is 18.15\$/Pixel. The cost of roads was obtained from the General Directorate of Transportation for the provincial type, which is 751,100 for one km and computed as 11.55\$/Pixel.

The risk was computed for each property (building and road) separately. Then the total risk map was obtained by overlaying all risk maps. The risk computation involves the multiplication of each property value with MDR values and then multiplication by hazard value.

$$\text{Risk}_{\text{Prop}} = R_{\text{Building}} + R_{\text{Road}} \quad (2)$$

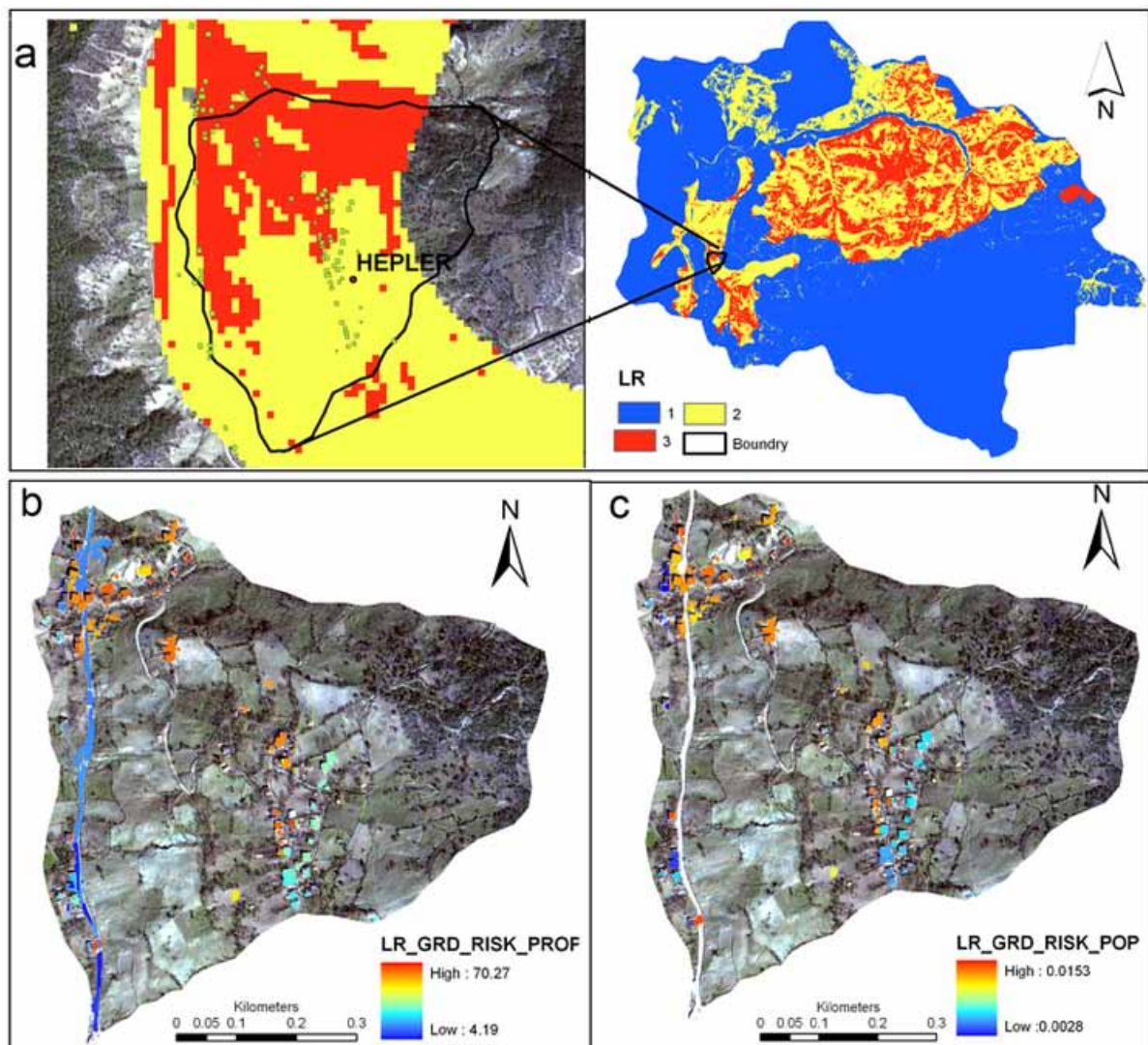
$$\text{Risk}_{\text{Prop}} = H \times (\text{MDR}_{\text{Building}} \times 18.15\$/\text{Pixel}) + H \times (\text{MDR}_{\text{Road}} \times 11.55\$/\text{Pixel}).$$

As a result, for the assumed damage probability matrix in Table 3, the annual risk to Hepler was estimated for TL/pixel cost of damage for grid and slope-unit based mapping units and for different mapping methods. Fig.6b illustrates the spatial variation of risk values for risk to the property.

The risk of loss of life can be computed by the sum of the multiplication of H and MDR values for damage states for loss of life as given in Eq (3):

$$R_{\text{LOL}} = H \left[P_{\text{HS}} \left(\text{MDR}_{\text{HSDT}} + \text{MDR}_{\text{HSNT}} \right) + P_{\text{LS}} \left(\text{MDR}_{\text{LSDT}} + \text{MDR}_{\text{LSNT}} \right) \right] \quad (3)$$

Figure 6. The grid-based LR hazard map levels for Hepler district (a); Risk to property (\$/Pixel) Max value: 54.053 (\$/Pixel), Min Value: 3.22 (\$/Pixel) (b); Risk of loss of life (Loss of Life/Pixel) (c)



As a result, the annual risk maps for loss of life were estimated, which is illustrated in Fig.6c.

FUTURE RESEARCH DIRECTIONS

Landslide risk assessment is still in its developing stage, and most countries do not have a standardized landslide risk assessment process. In Turkey, there are a few studies for risk mapping for landslides. Therefore, this study contributes to the literature by designing a loss estimation approach for quantitative risk assessment.

The developed algorithm to obtain features from high-resolution RS images provides an automatic way for obtaining elements at risk features. The cost of these images is comparatively lower for a small region than field observations when considering transportation and individual expenses in addition to

expenses for equipment. However, for future work, this algorithm can be improved to reduce errors by adapting new image processing methods suitable for rural areas. The use of the velocity value of run-out is new for vulnerability assessment and several issues such as rheology computation and model calibration are needed to be incorporated into the analysis. In addition to temporal impact, which is considered in the vulnerability approach, the probability of being in run-out zone for elements at risk, the construction type, the replacement cost of particular buildings, etc. may also be considered for further vulnerability assessment studies.

CONCLUSION AND DISCUSSION

The risk maps were obtained by the evaluation of hazard and consequence maps for each element at risk. A combined landslide risk map was obtained by adding risk for each element at risk. As a result, quantitative risk maps were produced on a continuous scale where numerical values indicate the distribution of risk including the annual probability of expected losses in \$ per pixel and the annual probability of life loss per pixel for property and life, respectively. The risk maps for loss of property provide a low risk value for the road; however, the risk values vary for building and provide medium and high levels of risk for buildings. Similar to loss of life, the buildings at the northern part exhibits higher risk values compared to the southern part of the region which might be due to the physical properties of the terrain which is related to the slope angle.

There is not a common method of vulnerability assessment in landslide risk mapping. The vulnerability approach designed for this study is based on the damage probability matrix and run-out velocity, which is novel to assess the vulnerability values quantitatively.

Consequence analysis, which involves elements at risk and vulnerability, is crucial for risk mapping. The elements at risk and vulnerability values depend very much on the scale of the investigation. Thus, the changes in consequence at different scales should be considered to assess changes in the resultant maps. The performance of risk mapping is strongly dependent on data constraints and financial supports. Additionally, each component of the risk assessment procedure requires the incorporation of data from diverse disciplines due to its multidisciplinary structure.

Therefore, the quality, reliability, completeness, and resolution of data should be considered prior to analysis.

The result of the analysis indicates that GIS and RS are very efficient technologies to handle, analyze, evaluate and overlay several data layers relevant to risk analysis. These maps are useful to evaluate the landslide risk of a region and it provides disaster mitigation benefits to housing and infrastructure. By using these maps, decision-makers can develop risk mitigation strategies and perform cost-benefit analyses for future land use planning.

ACKNOWLEDGMENT

This research was supported by ÖYP research foundation No: BAP-08-11-DPT.2002K120510 and partly supported by TÜBİTAK 1002 project No: 109Y003. Additionally, it is based on the Ph.D. study of the first authors.

REFERENCES

- Aleotti, P., & Chowdhury, R. (1999). Landslide hazard assessment: Summary review and new perspectives. *Bulletin of Engineering Geology and the Environment*, 58(1), 21–44. doi:10.1007/100640050066
- Amini, J., Lucas, C., Saradjian, M., Azizi, A., & Sadeghian, S. (2002). Fuzzy Logic System for Road Identification Using IKONOS Images. *Photogrammetric Record*, 17(99), 493–503.
- Aytekin, Ö., Erener, A., Ulusoy, İ., & Düzgün, Ş. (2012). Unsupervised building detection in complex urban environments from multispectral satellite imagery. *International Journal of Remote Sensing*. doi:10.1080/01431161.2011.606852
- Bell, R., & Glade, T. (2004). Quantitative risk analysis for landslides – Examples from Bildudalur, NW-Iceland. *Natural Hazards and Earth System Sciences*, 4(1), 117–131. doi:10.5194/nhess-4-117-2004
- Carrara, A., Cardinali, M., Guzzetti, F., & Reichenbach, P. (1995). GIS technology in mapping landslide hazard. In A. Carrara & F. Guzzetti (Eds.), *Geographical information systems in assessing natural hazards, advances in natural and technological hazards research* (Vol. 5, pp. 135–176). Springer. doi:10.1007/978-94-015-8404-3_8
- Chowdhury, R. N., & Flentje, P. N. (2002). Uncertainties in rainfall- induced landslide hazard. *Quarterly Journal of Engineering Geology and Hydrogeology*, 35(1), 61–70. doi:10.1144/qj.35.1.61
- Christophe, E., & Inglada, J. (2007). Robust Road Extraction For High Resolution Satellite Images. *ICIP 2007. IEEE International Conference on Image Processing*. 10.1109/ICIP.2007.4379859
- Chung, C. F., Fabbri, A. G., & Van Westen, C. J. (1995). Multivariate regression analysis for landslide hazard zonation. In A. Carrara & F. Guzzetti (Eds.), *Geographical Information Systems in Assessing Natural Hazards* (pp. 107–133). Kluwer. doi:10.1007/978-94-015-8404-3_7
- Chung, C. J. F., & Fabbri, A. G. (1999). Probabilistic Models for landslide Hazard Mapping. *Photogrammetric Engineering and Remote Sensing*, 65, 1389–1399.
- Dai, F. C., Lee, C. F., & Ngai, Y. Y. (2002). Landslide risk assessment and management: An overview. *Engineering Geology*, 64(1), 65–87. doi:10.1016/S0013-7952(01)00093-X
- Düzgün, H. S. B. (2008). A Quantitative Risk Assessment Framework for Rock Slides. *Proc. of 42nd US Rock Mechanics Symposium*.
- Erener, A. (2009). *An Approach for Landslide Risk Assessment by Using Geographic Information Systems (GIS) and Remote Sensing (RS)* (PhD Thesis). Middle East Technical University (METU).
- Erener, A., & Düzgün, H. S. B. (2009). Prediction of Population in Urban Areas by Using High Resolution Satellite Images. 4th international conference on Recent Advances in Space Technologies (RAST 2009). Space for the Developing World, İstanbul, Turkey.
- Erener, A., & Düzgün, H. S. B. (2010). Improvement of Statistical Landslide Susceptibility Mapping by Using Spatial and Global Regression Methods in the case of More and Romsdal (Norway). *Landslides*, 7(1), 55–68. doi:10.1007/10346-009-0188-x
- Erener, A., & Düzgün, H. S. B. (2012). Landslide Susceptibility Assessment: What are the Effects of Mapping Unit and Mapping Method? *Environmental Earth Sciences*. DOI doi:10.1007/12665-011-1297-0

- Erener, A., Mutlu, A., & Düzgün, S. H. (2016). A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM). *Engineering Geology*. doi:10.1016/j.enggeo.2015.09.007
- Erener, A., Sivas, A. A., Selcuk-Kestel, A. S., & Düzgün, H. S. (2017). Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. *Computers & Geosciences*, 104, 62–74. doi:10.1016/j.cageo.2017.03.022
- Fell, R. (1994). Landslide risk assessment and acceptable risk. *Canadian Geotechnical Journal*, 31(2), 261–272. doi:10.1139/t94-031
- Glade, T. (2003). Landslide occurrence as a response to land use change: A review of evidence from New Zealand. *Catena*, 51(3–4), 297–314. doi:10.1016/S0341-8162(02)00170-4
- Hidalgo, C., Vega, J., & Parra, M. (2018). Effect of the Rainfall Infiltration Processes on the Landslide Hazard Assessment of Unsaturated Soils in Tropical Mountainous Regions. Publicado en: Engineering and Mathematical Topics in Rainfall. *InTechOpen*.
- Jibson, R. W., Harp, E. L., & Michael, J. A. (1998). A method for producing digital probabilistic seismic landslide hazard maps: An example from the Los Angeles California area. *U.S. Geological Survey Open-File Report*, 98, 113–117.
- Kim, W.Z., & Nevatia, R. (1999). Uncertain Reasoning and Learning for Fearum Grouping. *Computer Vision and Image Understanding*, 76(3), 278–288.
- Klang, D. (1998). Automatic Detection of Changes in Road Databases Using Satellite Imagery. *International Archives of Photogrammetry and Remote Sensing*, 32(4/1), 293–298.
- Ko Ko, C., Flentje, P., & Chowdhury, R. (2004). Interpretation of Probability of Landsliding Triggered by Rainfall. *Landslides*, 1(4), 263–275. doi:10.1007/10346-004-0031-3
- Laptev, I., Mayer, H., Lindeberg, T., Eckstein, W., Steger, C., & Baumgartner, A. (2000). Automatic extraction of roads from aerial images based on scale space and snakes. *Machine Vision and Applications*, 12(1), 23–31. doi:10.1007/001380050121
- Lee, E. M., & Jones, D. K. C. (2004). *Landslide risk assessment*. Thomas Telford. doi:10.1680/lra.31715
- Leone, F., Aste, J. P., & Leroi, E. (1996). Vulnerability assessment of elements exposed to mass-movement: working toward a better risk perception. *Landslides*, 1, 263–269.
- Lin, C., & Nevatia, R. (1998). Building detection and description from a single intensity image. *Computer Vision and Image Understanding*, 72(2), 102–121. doi:10.1006/cviu.1998.0724
- Lu, D., Weng, Q., & Li, G. (2006). Residential population estimation using a remote sensing derived impervious surface approach. *International Journal of Remote Sensing*, 27(16), 3553–3570. doi:10.1080/01431160600617202
- Luna, B. Q., Blahut, J., Camera, C., Westen, C. V., Apuani, T., Jetten, V., & Sterlacchini, S. (2014). Physically based dynamic run-out modelling for quantitative debris flow risk assessment: A case study in Tresenda, northern Italy. *Environmental Earth Sciences*, 72, 645–661.

Mejia-Navarro, M., Wohl, E. E., & Oaks, S. D. (1994). Geological hazards, vulnerability and risk assessment using GIS: Model for Glenwood Springs, Colorado. *Geomorphology*, 10(1-4), 331–354. doi:10.1016/0169-555X(94)90024-8

Mena, J., & Malpica, J. (2003). Color Image Segmentation Using the Dempster-Shafer Theory of Evidence for the Fusion of Texture. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34(3/W8), 139–144.

Morton, T. A., & Yuan, F. (2008). Analysis of population dynamics using satellite remote sensing and US census data. *Geocarto International*, 2008, 1–20.

O'Brien, J. S. (1999). *Flo2D Users Manual Tetra Tech ISG*. Nutrioso.

O'Brien, J. S., Julien, P. Y., & Fullerton, W. T. (1993). Two-dimensional water flood and mudflow simulation. *Journal of Hydraulic Engineering*, 119(2), 244–261. doi:10.1061/(ASCE)0733-9429(1993)119:2(244)

Öztürk, K. (2002). Heyelanlar ve Türkiye'ye Etkileri, *G.Ü. Gazi Eğitim Fakültesi Dergisi*, 22(2), 35–50.

Peng, J., & Jin, Y. Q. (2007). An unbiased algorithm for detection of curvilinear structures in urban remote sensing images. *International Journal of Remote Sensing*, 28(23), 5377–5395. doi:10.1080/01431160601075574

Pereira, S., Garcia, R., Zêzere, J. L., Cruz, S. O., & Silva, M. (2017). Landslide quantitative risk analysis of buildings at the municipal scale based on a rainfall triggering scenario. *Geomatics, Natural Hazards & Risk*, 8(2), 624–648. doi:10.1080/19475705.2016.1250116

Rautela, P., & Lakhera, R. C. (2000). Landslide risk analysis between Giri and Tons rivers in Himachal Himalaya (India). *IAG*, 2(3/4), 153–160.

Remondo, J., Bonachea, J., & Cendrero, A. (2008). Quantitative landslide risk assessment and mapping on the basis of recent occurrences. *Geomorphology*, 94(3-4), 496–507. doi:10.1016/j.geomorph.2006.10.041

Silva, M., & Pereira, S. S. (2014). Assessment of physical vulnerability and potential losses of buildings due to shallow slides. *Natural Hazards*, 72(2), 1029–1050. doi:10.1007/11069-014-1052-4

Toprak Su. (1978). *Konya Kapalı Havzası Toprakları*. Toprak Su Genel Müdürlüğü Toprak Etüdleri ve Haritalama Dairesi Topak Etüdleri Fen Heyeti Md., Ankara. Yayın No: 288.

Uzielli, M., Düzgün, H. S. B., & Vangelsten, B. V. (2006). *A first –Order Second-Moment Framework for probabilistic estimation of vulnerability to landslides*. ECI Conference: Geohazards - Technical, Economical and Social Risk Evaluation, Lillehammer, Norway.

Vakalopoulou, M., Karantza, K., Komodakis, N., & Paragios, N. (2015). Building Detection In Very High Resolution Multispectral Data With Deep Learning Features. In *IGARSS 2015*. IEEE.

Vega, J., & Hidalgo, C. (2017). Risk Assessment of Earthquake-Induced Landslides in Urban Zones, in *Advancing Culture of Living with Landslide – Advances in Landslide Science* (Vol. 2). Springer International Publishing AG, 2017, 953–963.

Vega, J. A., Marín, N. J., & Hidalgo, C. A. (2019). Statistical Approaches for The Assessment of Landslide-Related Economic Losses. In *IOP Conf. Series: Materials Science and Engineering*. IOP Publishing. doi:10.1088/1757-899X/471/10/10200

Wang, S. Q., & Unwin, D. J. (1992). Modelling landslide distribution on loess soils in China: An investigation. *International Journal of Geographical Information Systems*, 6(5), 391–405. doi:10.1080/02693799208901922

Yang, J., & Wang, R. S. (2007). Classified road detection from satellite images based on perceptual organization. *International Journal of Remote Sensing*, 28(20), 4653–4669. doi:10.1080/01431160701250382

Zhang, B. G. (2003). Application of Remote Sensing Echnology to Population Estimation. *Chinese Geographical Science*, 13(3), 267–271. doi:10.100711769-003-0029-0

ADDITIONAL READING

Erener, A., & Düzgün, H. S. B. (2010). Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway). *Landslides*, 7(1), 55–68. doi:10.100710346-009-0188-x

Erener, A., & Düzgün, H. S. B. (2012). Landslide susceptibility assessment: What are the effects of mapping unit and mapping method? *Environmental Earth Sciences*, 66(3), 859–877. doi:10.100712665-011-1297-0

Erener, A., & Düzgün, H. S. B. (2013). A regional scale quantitative risk assessment for landslides: Case of Kumluca watershed in Bartın, Turkey. *Landslides*, 10(1), 55–73. doi:10.100710346-012-0317-9

KEY TERMS AND DEFINITIONS

Element at Risk: Features under risk including buildings, roads, land use, and infrastructure.

GIS: Geographic information systems is a technology that is used to capture, store, manipulate, analyze, manage, and present spatial data.

Landslide Hazard: Information about probability of occurrence of a landslide in a given area over a specified period of time.

Landslide Risk Assessment: Establishing the likelihood to which future slides could adversely impact society.

RS: Remote sensing is the acquisition of information from earth without making physical contact with the object by using satellite- or aircraft-based sensor technologies.

Vulnerability: Defined as degree of loss, suffer harm for a given element or set of elements within the area of hazard event.

Section 13

Human Aspects of Technology

Technology and the Theory of Apocalypse

Maximiliano Emanuel Korstanje

 <https://orcid.org/0000-0002-5149-1669>

University of Palermo, Argentina

INTRODUCTION

The television series *The Walking Dead* has reached more than seven million “likes” on Facebook and more than 300.000 in Twitter through March of 2012 (March, 2012), which lead very well scholars to think that the popular culture has widely adopted the figure of Zombie as a new genre of entertainment. The idea of the “Undead” not only coincides with the contemporary needs of morbid consumption (gaze) but also with the rise of new capitalism that commoditizes “the other’s death as a vehicle towards the individual emancipation (Korstanje 2012). The undead, broadly speaking, inhabit and proliferate in this earth while they are undesired by the human civilization. They were like us, but today they are death!. The Walking dead appeals to a climate of total destruction which often does not resonate with a form of public consciousness, but in the collective unconsciousness. The undead represents a postmodern sensibility. This sensibility reeks of decay.

“[It is a] ‘degraded’ landscape of schlock and kitsch, of TV series and Readers’ Digest culture, of advertising and motels, of the late show and the grade-B Hollywood film, of so-called para literature with its airport paperback categories of the gothic and the romance, the popular biography, the murder mystery and science-fiction or fantasy novel: materials they no longer simply ‘quote’, as a Joyce or a Mahler might have done, but incorporate into their very substance”. (Jameson 1991:55)

As the previous backdrop, Jameson argues that the postmodern culture rests on representations revolving around the “unconscious”, in which case, it articulates a specific discourse that interrogates in the political sphere. In a similar vein, the arts have represented, in what might be considered prescient ways, what has already begun in the economic, political, and social structures, but has not yet appeared in explicit terms. Somewhat arbitrarily stated, modernism began in the mid-nineteenth century. Charles Baudelaire called it *modernité*.

As Ulrich Beck puts it, the obsession of modern man to control the nature opened the doors to the idea that the same technology used by making of this world a better and safer place, can usher mankind into a nuclear apocalypse. One of the paradoxes of modernity began but it was not limited to the Chernobyl Accident. Technology played a leading role in the configuration of a Cold War, where the US and the Soviet Union devoted their resources to an incredible arms race. Beck acknowledges that at the time some risks are successfully mitigated by the action of technology and rational technique, new ones emerge. What is more important, the state of disaster results from the excess of modernity (Beck 1992). Jacques Ellul interrogated philosophically on the negative effects of technology in modern society, not only undermining critical thinking but also controlling -if not commoditizing- other lay-citizens (Ellul 1962; 1992). Andrew Feenberg calls the attention on technology, a-la-Marcuse, as a form of new rationalization, which means control over the rank-and-file worker. Technology disposes not only from bodies

DOI: 10.4018/978-1-7998-3479-3.ch112

to enhance the means of production but tries to domesticate nature through the capitalist wage system. While machine operates to make our lives safer, the western rationality imposes over other voices and cosmologies (Feenberg, 1995).

In the present chapter, we interrogate not only on the nature of technology but its connection with the bottom days theory and the Zombie world. After reviewing plots as *The Matrix*, *Noah and the night of the Living Dead* (among others), we hold the thesis that technology gives humans the opportunity to emancipate themselves from their limited condition, but at the same time, it may be very well a potential oppressor when ethics is derided. The chapter is not intended originally to managers and policymakers, simply because I am a social scientist. However, top-managers can enrich their diagnosis of organizational issues after reading it. Another limitation of the work consists in its lack of empirical evidence. Whatever the case may be, it provides fresh anthropological insight about the role of technology in the social imagination, to be more exact in the archetype of the bottom days. From the immemorial times, technology prompted the rise of empires and great civilizations while the precondition for a climate of extreme violence or war that accelerated the decline of society. The question whether the sin, greed, and omnipotence ushered mankind into the ruin was widely discussed within historians. This chapter explores the theory of apocalypse as a precondition where humans are not destroyed but purged towards a new *rebirth*.

Background Section

Every culture has tales about an exemplary center, an immemorial life where human beings and god coexisted in harmony. This paradise, a term coined by Eastern religions, has prevented access by humans the founding parents committed a crime, sin or other norm violation. From that day on, civilizations try to regain the primitive state of nature where suffering did not take place (Cohn 1996; McGinn 1996; Kumar 1996). One of the intriguing aspects that characterize the end of the World associates to what Eric Ranking dubbed as “the introduction of dark technology”. To put this in other terms, he holds the thesis that the world, no matter the tale, will never be destroyed but it is only reconstructed by the articulation of a purge. Humans not only offended the Gods following their greed but also are a real threat to the natural order. Gods conduct a purge to correct this “original sin” while the earth is redeemed into a new sanitized facet of existence. In these terms, as Rabkin explains, the exile represents the human impossibility to return to the exemplary center, but what is more important, it reminds the world may be destroyed because of human greed (Rabkin 1986). When the world is destroyed, what humans do not disappear, but they are forced to live in a new condition. In consonance with Rabkin, Gary Wolfe uses the term *Zero Moment* to refer the return to a primitive state where humans struggled for surviving. Once the world is destroyed, simply the hierarchal authorities are withered away, as well as all constitutional rights. While a new aristocracy arises, the state of prosperity or well-being earlier the moment of Armageddon never comes back. Technology paves the ways for the emergence of self-contained empires, while gradually when it is instrumentalized to protect the interests of a ruling elite, corruption marks the beginning of the end of the human race. The term, *remaking zero* crystalizes the efforts of humans to understand their being in this world. The civilization is born and dies thanks to technology, reminding that.

“The promise inherent in the idea of remaking zero is certainly one of the reasons this genre has survived as long as it has, and in so many guises. On the simple level of narrative action, the prospect of a depopulated world in which humanity is reduced to a more elemental struggle with nature provides a

convenient arena for the sort of heroic action that is constrained in the corporate, technological world that we know” (Wolfe, 4).

In this respect, *The Matrix* offers a good explanation that ascertains how and why technology is a key-player of human decline. Let's firstly clarify this Saga, which was starred by Keanu Reeves, Carrie-Anne Moss and Laurence Fishburne narrate the grim days of humanity which is exploited by Machines. Needless to say, the plot situates us in a futurist world, where humans have been enslaved by “the Matrix”, a great computer dotted of artificial intelligence. The Matrix -though created to serve humans- changes its orders subduing them into a fictional (simulated) reality. In the simulation, people are free to travel, work and live a normal life, but they are connected to a pod that provides to the Matrix of the necessary electrical activity the body offers. A young computer programmer (Thomas Anderson), Neo discovers that he was living in a simulation. He believes something is not working fine but continues without knowing he will be part of something great. Anderson is met by Trinity (Moss), a hacker who introduces Morpheus (Fishburne) to Neo, who is invited to make a difficult decision. On his hands, there are two pills. The red pill offers a revelation but leading Neo towards a hostile (but real) world. Rather, the blue pill leaves things as they stand ushering Neo into a comfortable but routinized world. Neo swallows the red option disintegrating him and appearing naked in a liquid-filled pod where he is rescued by Trinity and Morpheus. The plot is that in the 21st-century machines are empowered by humans to embrace artificial intelligence. In a rebellion, these smart machines confront humanity in a bloody war. At the time, the sun was blocked to prevent their advances so the machines cultivated human bodies as valid sources of energy, as a harvest disposed to extract bioelectricity. In one of its adventures, Neo and his friends enter the Oracle who prophesized that Neo is the One, a type of chosen messiah that will liberate humans from the machine's yoke. Morpheus and Trinity are betrayed by Cypher a crew member who renounces to the sense of reality for the wonderland offered by the matrix. Cypher negotiates with Agent Smith in exchange for having a comfortable life. Cypher disconnected some of the crew members while they were in the matrix which meant certain death, but failed in doing the same for Trinity and Morpheus. Cypher is killed by Tank. Morpheus is interrogated to know the exact location of Zion, the last rebel base under the earth. Trinity and Neo help Morpheus in escaping from Smith's traps. Here we have shortly enunciated the main points of the plot which in recent years have captivated the attention of writers, thinkers, scholars and philosophers. To some extent, the *Matrix* reflects the philosophical discussion on reality and pseudo-reality or what was Morpheus's phrase “welcome to the desert of the real”. Professor William Irwin (2005) edited a book entitled *The Matrix and Philosophy* where he interrogates on the parallel between Neo and Socrates. Anderson wakes up in a reality he never thought about and he renounces to the comfortable wonderland he lived only by grasping the truth (Irwin, 2005). The spirit of the *Matrix* reveals two important things to our argumentation. On one hand, from Cartesian dualism onwards, philosophers have indagated on the role of scepticism to lay the foundations of scientific inquiry. The sense of truth can be tainted not only by mental illness but also by dreams which are “an uncanny state of the mind”, where reality is depicted beyond the scrutiny of the dreamer and their will. Although we may very well perceive the world around us in a dream as real, it is unreal. It is flourishing from our inner-life toward a mirror-like image. Of course, the sense of living, in reality, involves us entering into an existential pain, which not all can bear. This is exactly the dilemma in Cypher, who adopts a better life instead of a “real life”. This suggests that what is defined as “the real life” allows us to make decisions on a freeway, although our basic resources for survival are not granted or our well-being jeopardized (Irwin 2005). In consonance with this, David M. Nixon (2005) argues that believing is an act strongly associated to previous experiences in a way that sometimes leads us to

simulation. Neo not only believed he was living a real life, oddly with his experiences being fabricated by an evil machine (the Matrix) but paradoxically, he believes in Morpheus when he asks for knowing more about the real story. In this point, Carolyn Korsmeyer (2005) presents an interesting thesis which merits discussion. In her view, the *Matrix* triggers two different questions; one consists of asking what is real and valuable? and the second is, is what is real that which gives a more pleasurable experience?

The *Matrix* reminds us that what we see can be *the illusory product of a program*, but it paves the way for the rise of anti-Western rationalism that contradicts the hegemony of gazing as the best method to reach the truth. Western civilization has disposed of the *gaze as a metaphor for understanding* but things move from worse to worst, and the mind maddens if reality is perceived as it is. The betrayal of Cypher explains the urgency of forgetting the past and prioritizing the pleasure of the body over existential interrogation.

In a nutshell, Matrix puts in debate two important aspects of apocalypse. On one hand, the apocalypse should be understood as the rise of reality into the fictional world of immediate gratification, which is often constructed by humans. On another, sooner or later, this fake world melts in the air in the same way that the principle of reality takes from the slumber we are. This point was brilliantly assessed by Lord Raglan, who was one of the pioneers in the study of mythical heroes. Like Neo, the mythical heroes are blind to their promising future, and they were persecuted from their childbirth, even educated by other families abroad. Their royal lineage placed them in a real risk of being assassinated by the ruling monarch. In their mythical adventures, heroes face major threats and countless obstacles, which probes the human character (Raglan 2013). To some extent, Raglan amply shows how the archetype of the mythical hero fits with the needs of putting humankind into a test. This is the main discourse enrooted in the core of the theory of apocalypse.

As the previous argument is given, Wolfe enumerates some specific feats that the bottom-days narratives share such as the experience of the cataclysm, a journey through the wasteland created by the disaster, the settlement of a new community and the decisive battle to impose human values. A whole portion of novels reveal that the central character of the hero remains isolated when the disaster takes hit. This isolation was in view of an original fault, or simply as a moral punishment for an involuntarily crime. Almost all heroes are marked (if not cursed) by a sentiment of loneliness which ushers them in a depression or madness. The hero should travel far to understand the magnitude of cataclysm or simply due to a moral purgation for its personal offenses. This point is vital for a new community to be settled once hero defeats the forces of evilness reminding one more time why humans are so special. To cut the long story short, let's add that horror movies and apocalyptic genre were historically intertwined. As we shall see in the next, the archetype of heroism, as well as the hero's character, prevail in almost all plots.

The Zombies in Horror Movies

George Romero was the first film-maker who introduced the figure of Zombies in the horror movies. In the night of the Living Dead (1968) or Day of the Dead (1985), the intriguing double-edge of technology occupies a central role. For some reason, which is not specified in the plot, flesh-eating zombies populate the world pressing mankind to the extermination. The numbers of zombies outnumber humans 400.000 to 1. Dr Sarah Bowman and Miguel Salazar escape to Fort Myers, Florida where they (jointly others as Helicopter pilot John, or Bill). However things come worse to worst, and they are pressed to make way towards an army base located in Everglades. Once there, they are hosted by Captain Rodhes and Dr. Logan, who commands a group of scientists and militaries whose surgical experiment with Zombies. At this lab, these survivors devote considerable efforts and resources to domesticate (civilize)

the cannibalistic zombies. Sooner the tension between soldiers and scientists is evident and Rhodes ask their troops to destroy all the experiments in the lab. In an instance, Johnny -the pilot- invites Sarah to drink and relax while Sarah demands him to be unwilling to subscribe to neither side. He replies to her, that the world is exhausted humans should face the punishment of God because of their greed. Most likely, Johnny bets, the Lord imposed the mankind extermination to protect the earth he does not want to see how humans eagerly devour the natural resources. Beyond the extreme suffering, this evokes a renovation for the human race, and of course for the planet. The same can be found in *The Resident Evil 2*, where Umbrella Corporation manipulates a lethal virus that rapidly expands worldwide. The protocols of contention failed and sooner the virus widespread everywhere outside The Hive, the underground laboratory where everything started. Alice and friends struggle against this corporation without knowing she has genetically manipulated to be the “perfect warrior”. There are many commonalities between the two plots, which merit to be discussed. On one hand, the use and abuse of technology imply a great risk for mankind. The cannibal “living dead” return to life to eat human flesh while a bunch of survivors struggles to look another safer place to settle. The forces of evil and good enter in the last of battles, where the values that will reign on the earth are at stake. This mythical conflagration is accompanied by the needs of expiating the old human sins. Essential interests are the different matter in the plot of *Contagion* which emulates the latest Swine Flu (H1N1) outbreak in 2009-2010. After a business trip bound for China, Beth Emhoff (Gwyneth Paltrow) return to Chicago, US to have an adventure with her former lover earlier than going to her family in Minneapolis. She is married to Mitch (Matt Damon) who is not aware of her infidelity. The condition of Beth is suddenly worsened and it finally aggravates her death. Her six-year-old son dies as well. Though Mitch is placed in isolation, he and his daughter seem to be immune. Meanwhile, in Atlanta authorities and scientists of the Department of Homeland Security have a bolster with Dr. Ellis Cheever (Lawrence Fishburn) to coordinate future efforts to contain and respond to a lethal pandemic. Once again, the sense of fault is remembered according to the origin and rise of the apocalypse. *Contagion*, unlike *Resident Evil*, signals to scientists –not as the evildoers- but as the heroes who ultimately found and implemented a vaccine that saves millions of lives. The lack of contact between infected and healthy persons as well as the role of mobilities in the configuration of nation-state appears to be some of the topics this film deals.

Last but not least, another interesting movie where the archetype of bottom-days can be debated is *Noah*, a film starred Russell Crowe (Noah) and Jennifer Connelly (Naameh). At a first glimpse, the plot is developed in a grey post-apocalyptic landscape where *The Watchers* are punished by God once Semyaza led them to procreate with women. Their offspring, the Nephilim devoured the cattle and were prone to devastate the earth. God disposed of a deluge which purged the earth while Semyaza and his fallen angels were castigated. Aronofsky, Noah’s director, combines the *Chronicles of Enoc* book (see on *The Watchers*), with the classic biblical texts to offer a more than the interesting product. In this plot, the Watchers not only helps Noah to struggle against *Tubal-Cain* but they are the conduit towards the Lords plan to succeed. God asked Noah to fabricate an Ark, but he is decisively encouraged not to tell the god’s plan to Cain and his offspring. Noah should select a couple for species to continue with life. The ark contains not only the germen of life in the next world, prepared by God but also the future of humanity in the hands of Noah. The problem is that Noah is urged to keep the secret, witnessing how the human is ruthlessly obliterated through the articulation of a universal flood. This extinction became in the first genocide and Noah the architect of a new world where the sins of man are redeemed. The “shameful” Noah falls into depression and drinks all the time until God blesses the family as the new start of a re-born human race. Well, to here, we tried to remark the main common-thread arguments in four movies, which were broadcasted in different cultural background and times. Here two interesting

point should be outlined. First and foremost, as I noted in *The Rise of Thana Capitalism and Tourism* (Korstanje 2016), Noah was the biblical myth that not only influenced modern audiences but also remained to the passing of time. Noah's ark likely revalidates the assumptions that life needs from death. Because of this, Thana Capitalism rests on the dynamics of creative destruction in the same way Noah is the architect of the first mass cleansing. The drives of death, which today are canalized through dark consumption and dark tourism, are legitimated by this biblical narration. God disposes to exterminate humankind but in secrecy situating Noah as a privileged witness of his plans. Secondly, what lies beyond the message of this myth is the ongoing competence of liberal market, which places the rank-and-file worker inegalitarian conditions to others –who like him- will fight to impose his position. Unlike other productive systems, where human ties were stable and durable, capitalism gradually erodes the trust liberating the human relations to capital owners. This led to precaritize the workforce dealing with new emerging institutions. In the days Thana-capitalism, citizens reinforce their loyalties in their institutions and their auratic sentiment of supremacy by gazing at how others die. In TV programs, Magazines, Press, Realities, Novels, Movies and other sources of media entertainment, (even dark tourism), the others' death is systematically framed, packaged and distributed as the main commodity to exchange. We are obsessed to consume the suffering while by this rite we are safe at home. This happens because of two main reasons. As discussed in earlier sections, the process of secularization not only undermines the influence of religion and the belief in the after-life. Death and the act of dying to begin to be conceived as a sign of inferiority. It is important not to lose the sight of the fact the global audiences undergird their sentiment of supremacy while gazing (consuming) news of disasters, mass death, cleansing and other calamities. They feel part of the “chosen peoples” who are immune to the cataclysm. We toy with the belief this sentiment builds a strong narcissism, which unless regulated, may lead towards “chauvinist” reactions. We put the example of *Hunger Games*, Collins' novel, where each participant (tribute) keeps an overrated-image of itself while at the bottom all will suffer the same end. The rebellion only started when involved actors opt to cooperate with the other. The same applies to *Thana-Capitalism*, where the worker is pitted against the worker in a marketplace where few hands concentrate a whole portion of yielded wealth while a whole life with nothing. Last but not least, Narcissism, which is stimulated by hyper-consumption and aesthetics, is vital to understand the hegemony of the ruling elite. The meaning of apocalypse has a double-edged function. While lay-citizens valorize technology as the touchstone of modern society, its negative effects are not discussed. E. Snowden not only showed how a democratic government would vulnerate the individual rights to privacy, but also the uses of surveillance high tech when the ideals of democracy are subordinated to fear (Altheide, 2014; Skoll, 2014; Korstanje & Skoll 2018).

Future Research Direction

As declared, this paper is not originally aimed at triggering a debate to managerial or administrative issues. Rather, it focuses strongly on the role played by technology in the configuration of the end of the world. But what is the connection of this mythology with the economic world?

To some extent, our means of production, which are enrooted in the economic matrix that regulates society, are drawn by our myths and myth-formation. Far from being naïve stories told and retold by ancient tribes, myths represent real symbolic guidelines which help modern citizens to resolve their problems according to how their founding parents did. What is more important, myths are culturally created and circulated to resolve tensional problems or philosophical quandaries people are afraid such as the live, or death. The idea beyond the end of the world, as Eliade puts it, coincides with our own

preconceptions which led historically to organize labor in cycles of production and destruction. Basically, the use of technology tends to modify the territory but in so doing, humans need to deal with a sentiment of culprit. As administrators of the world, humans employ an abstract rationality to place order in the animal kingdom. However, the introduced technology not only alters the environmental condition of life, subordinating the natural world to the culture, but also creates a sentiment of self-punishment which is symbolically expressed in the belief the world should suddenly end. The question of whether the rationality exhibits an original sin which needs to be redeemed is not new, and of course it is not a monopoly of Abrahamic Faiths. Men are outstanding creatures enthralled to administer the nature, but in some extent they fail, and for that they should be punished. Last but not least, the reconstruction of human life, just after a disaster takes hit, has been commoditized and formed as a cultural spectacle which today captivates global audiences worldwide. The horror-movie industry just in post 9/11 days has appealed to the theory of the end of the world to domesticate the discontents and frustrations in the workforce.

CONCLUSION

After a further review, this chapter ignited an interesting discussion around the idea of the apocalypse or at the best to the role of technology by slicing humanity from nature. In terms of the British anthropologist Tim Ingold (2000), one of the conceptual limitations of “dwelling perspective” exhibits the clear-cut division of humans from the natural environment. Unlike hunters and gatherers who have developed “relational” ties with the sensual world, we are educated to imagine ourselves as administrators of the natural world. Because of this, the eco-friendly projects (as conservation parks) often exclude the presence of humans. The employment of technology denotes certain rationality, which while sorting the environment according to our needs, creates a sentiment of the culprit. The theory of apocalypse expresses the “original sin”, the sentiment of culprit which places human values in jeopardy. This happens because mankind, as an outstanding creature, is the only one gifted by Gods to administer nature, but inevitably failed. Moved by greed and speculations, men governed with energy in backs to God and for that, they should be heavily punished. As a result of this, the diaspora of survivors appears to be in quest of a new settling where humans give rise to the new cities, a type of rebirth that alternates the fear of coming back to the nightmare of past and the hope for a better future.

REFERENCES

AdditionalREADING

- Altheide, D. L. (2014). The Triumph of Fear: Connecting the Dots about Whistleblowers and Surveillance. *International Journal of Cyber Warfare & Terrorism*, 4(1), 1–7. doi:10.4018/ijcwt.2014010101
- Baldwin, G. R. (2007). World War Z and the end of religion as we know it. *Cross Currents*, 412–425.
- Beck, U. (1992). *Risk society: Towards a new modernity* (Vol. 17). Sage.
- Bishop, K. (2009). Dead man still walking. *The Journal of Popular Film and Television*, 37(1), 16–25. doi:10.3200/JPFT.37.1.16-25
- Cohn, N. (1996). Upon whom the ends of the ages have come. In M. Bull (Ed.), *Apocalypse Theory and the end of the World* (pp. 33–49). Blackwell.

- Comaroff, J., & Comaroff, J. L. (2002). Alien-nation: Zombies, immigrants, and millennial capitalism. *The South Atlantic Quarterly*, 101(4), 779–805. doi:10.1215/00382876-101-4-779 PMID:17494223
- Contagion. (2011). Steven Soderbergh. 106 minutes. Participant Media, US.
- Darlington, P. J. (1965). Bio-geography of the southern end of the world. *Bio-geography of the southern end of the world*.
- Drezner, D. W. (2014). *Theories of International Politics and Zombies: Revived Edition*. Princeton University Press. doi:10.2307/j.ctt7zvcg0
- Ellul, J. (1962). The technological order. *Technology and Culture*, 3(4), 394–421. doi:10.2307/3100993
- Ellul, J. (1992). Technology and democracy. In *Democracy in a technological society* (pp. 35–50). Springer. doi:10.1007/978-94-017-1219-4_3
- Feenberg, A. (1995). Subversive Rationalization: Technology, Power and Democracy. In *Thechnology and the politics of Knowledge*. Indiana University Press.
- Freedman, R. (2013). *The War to End All Wars: World War I*. Houghton Mifflin Harcourt.
- Gasparini, S. (2015). *Zombis, fantasmas y la representación de la violencia en la narrativa argentina reciente*. XXVII Jornadas de Investigación del ILH.
- Graphs.net. (2013). *Vampires on Social Media*. Accessed March 8 <http://www.graphs.net/201301/vampires-on-social-media.html>
- Hassler-Forest, D. (2012). Cowboys and zombies: Destabilizing patriarchal discourse in The Walking Dead. *Studies in Comics*, 2(2), 339–355. doi:10.1386/tic.2.2.339_1
- Ingold, T. (2000). *The perception of the environment: essays on livelihood, dwelling and skill*. Psychology Press.
- Irwin, W. (Ed.). (2005). *More Matrix and philosophy: Revolutions and Reloaded decoded*. La Salle, Open Court Publishing.
- Jameson, F. (1991). *Postmodernism, or the Cultural Logic of Late Capitalism*. Duke University Press.
- Keller, C. (2004). *Apocalypse now and then: A feminist guide to the end of the world*. Fortress Press.
- Korsmeyer, C. (2005). Seeing, believing, touching, truth. In W. Irwin (Ed.), *The Matrix And Philosophy* (pp. 41–52). Open Courts.
- Korstanje, M. E. (2009). El culto a los muertos: la mitología nórdica antigua en el cine moderno. *Revista Chilena de antropología visual*, 13, 61-78.
- Korstanje, M. E. (2012). *The Rise of Thanat Capitalism and Tourism*. Routledge.
- Korstanje, M. E. (2013). Tourism. *Mobility & Zombie War. IJSTH*, 1(4), 51–55.
- Korstanje, M. E., & Skoll, G. (2018). Technology and Terror. In *Encyclopedia of Information Science and Technology* (4th ed., pp. 3637–3653). IGI Global. doi:10.4018/978-1-5225-2255-3.ch316
- Kumar, K. (1996). Apocalypse, millennium and utopia today. In *Apocalypse Theory and the end of the World*. Blackwell.

Lazar, S. (2012). How Zombies Became a Social-Media Firebomb. *Entrepreneur*. <https://www.entrepreneur.com/blog/223075>

Lyons, J. C. (2009). *Perception and basic beliefs: Zombies, modules and the problem of the external world*. OUP USA. doi:10.1093/acprof:oso/9780195373578.001.0001

McGinn, B. (1996). the end of the world and the beginning of Christendom. In *Apocalypse Theory and the end of the -World* (pp. 75–108). Blackwell.

Morton, T. (2013). *Hyperobjects: Philosophy and Ecology after the End of the World*. University of Minnesota Press.

Night of the Living Dead. (1968). *George Romero (dir)*. 96 minutes. Image Ten.

Nixon, D. M. (2005). The Matrix possibility. In W. Irwin (Ed.), *The Matrix And Philosophy* (pp. 28–40). Open Courts.

Noah. (2014). Dalen Aronosfky (dir). 138 minutes. Regency Enterprise, US.

Rabkin, E. (1986). Introduction: why destroy the World. In *The End of the World*. Southern Illinois University Press.

Resident Evil 2. Apocalypse. (2004). Alexander Witt (dir). 94 minutes. Constantin Film, US.

Rousseau, J. J. (1997). *Rousseau: 'The Social Contract' and Other Later Political Writings*. Cambridge University Press.

Skoll, G., & Korstanje, M. (2014). The Walking Dead and Bottom days. *Antrocom: Online. Journal of Anthropology*, 10(1), 11–23.

Skoll, G. R. (2014). Stealing Consciousness: Using Cybernetics for Controlling Populations. *International Journal of Cyber Warfare & Terrorism*, 4(1), 27–35. doi:10.4018/ijcwt.2014010104

Solow, R. M. (1973). Is the End of the World at Hand? *The Challenge (Karachi)*, 16(1), 39–50. doi:10.1080/05775132.1973.11469961

The Day of the Dead. (1985). *George Romero (dir)*. 100 minutes. Dead Films.

Thompson, D. (1998). *The end of time: faith and fear in the shadow of the millennium*. University Press of New England.

Wallerstein, I. M. (1999). *The end of the world as we know it: Social science for the twenty-first century*. U of Minnesota Press.

Williams, E. C. (2011). *Combined and Uneven Apocalypse: Luciferian Marxism*. John Hunt Publishing.

Wojcik, D. (1997). *The end of the world as we know it: Faith, fatalism, and apocalypse in America*. NYU Press.

Wolfe, G. K. (1983). The Remaking of Zero: the beginning of the end. In *The End of the World*. Southern Illinois University Press.

KEY TERMS AND DEFINITIONS

13

Capitalism: This is an economic system which rests on the principle of private ownership and the use of means of production towards profit.

Precaritized Labour: Low quality jobs offered in decentralized economies.

Technology: This represents the set of techniques, knowledge, skills, and actions oriented to transform the environment.

Thana-Capitalism: This signals to a new stage of capitalism where the gazing of Other's death is the main commodity to exchange. The term was originally coined by Maximiliano E Korstanje in his book *The Rise of Thana Capitalism and Tourism*.

Theory of Apocalypse: It is part of scatology, a subfield of mythology that explains and vaticinates how the world will end.

Software Literacy as a Vital Digital Literacy in a Software–Saturated World

Craig Hight

University of Newcastle, Australia

Elaine Khoo

 <https://orcid.org/0000-0002-4260-4553>

The University of Waikato, New Zealand

INTRODUCTION

Software mediates almost every aspect of everyday life. Nearly all of our professional and personal activities are embedded and shaped within systems and interactions that involve software at some level; from the seductive ecosystems on devices such as smartphones or tablets, to the tools underlying everyday practices such as word processing and networked communication, through to the more functionally limited interfaces increasingly embedded into our homes and work spaces. None of these tools are ‘neutral’. They embody social and cultural assumptions about their use and all have particular values embedded in their interfaces and affordances. These reflect conceptual frameworks in which particular ways of acting or thinking are more possible and imaginable than others. Despite their ubiquity within contemporary society, software has only recently emerged as a field of study in its own right. As championed by the proponents of Software Studies this is a vital but neglected area of cultural production intersecting with and potentially shaping all other spheres of social, cultural, economic and political activity (Fuller, 2008; Johnson, 1997; Manovich, 2013). This article proposes the notion of ‘software literacy’ to highlight a neglected aspect of digital literacies, and a key means of conceptualizing the skills and understandings needed for people to be critical and creative users of software in today’s software-saturated culture. This contribution argues for the relevancy of software literacy as deeply intertwined with people’s engagement with software and how it influences the way people come to understand, represent, generate and critique knowledge.

The discussion below begins by overviewing the background and need for a focus on software studies before proceeding to define, introduce and elaborate on a framework for software literacy. An outline of the framework is then grounded in and exemplified through a case study located within a university teaching and learning context with implications for further thinking and research in the field.

BACKGROUND

The proliferation of digital and networked technologies is an expanding and accelerating feature of modern societies and can be predicted to continue to rise and impact on almost every sphere of human living. This trend is reiterated in UNESCO’s latest policy document calling for all countries to mobilize the potential of digital including mobile technologies “to strengthen education systems, knowledge dissemination, information access, quality and effective learning, and more effective service provision” if

DOI: 10.4018/978-1-7998-3479-3.ch113

they are to achieve the aspirational goal of inclusive and equitable quality education and lifelong learning by 2030 (UNESCO, 2015, p.37). Embedded within any digital device, software operates at multiple levels in contemporary life; as applications and platforms which we engage to facilitate and augment a host of social, economic and political practices, and less obviously as part of the taken for granted infrastructure of globalized cultural and economic exchange. We argue that software remains a conceptually neglected component of digital literacies, an often taken for granted aspect of the digital world which needs to be recognized as a distinctive part of contemporary literacies with its own dynamics and qualities which enable and constrain other literacies. Many practices within contemporary life, including those aggregated under the umbrella of digital, media and information practices are now ‘coded’ in the sense that they only exist and are constituted through programming code (Kitchin & Dodge, 2011).

However debates over terms such as information and communications technology (ICT) literacy, information literacy, digital literacy, media and information literacy, digital information literacy, and media and literacies studies tend not to acknowledge software as an actant in the world (Kitchin & Dodge, 2011) nor the emergence of software as a distinctive component of digital, media and information practices (Khoo, Hight, Cowie & Torrens, 2017). Our framework, software literacy, is intended to contribute to existing bodies of work on digital literacy (see for example, Ala-Mutka, 2011; Alexander, et al., 2017; Beetham & Sharpe, 2010; Hegarty et al., 2010; JISC 2014; Sefton-Green, Nixon, & Erstad, 2009; UNESCO, 2013). No studies to date that we know of raise the role of student understanding (including those who are new to using software) of how software and its affordances influence knowledge representation, generation and critique. Many studies have been conducted on information literacy and on ways of mastering software (Underwood, 2009), but the role of software itself tend to be taken for granted and is only now emerging as a field of study (Fuller, 2008; Lynch, 2015; Manovich, 2013).

Most people develop some level of proficiency with everyday software and hardware informally through their daily use and incremental engagement over time (Bulfin & Koutsogiannis, 2012; Hague & Logan, 2009). In educational research, informal learning practices have been shown to increase students’ sense of agency and consequently to have the potential to make learning a richer and more fulfilling experience (Furlong & Davies, 2012). Commenting on the trend of digital penetration, numerous authors have further argued that ubiquitous access to digital technologies has shaped a new internet-centred generation of ‘digital natives’ (Oblinger, 2003) with the corresponding assumption that access to digital tools has, on its own, facilitated the development of new learning skill sets (Tapscott, 2009). Terms such as the ‘digital generation’, ‘millennials’, ‘Net Generation’ (Tapscott, 1999), ‘digital natives’ (Prensky, 2001), ‘Google generation’, ‘Generation Y’ and so forth have derived from a host of assumption about the distinctive skill set of generations immersed within digital technologies. Such labels aim to characterize an emerging class of learners accustomed to engaging with software and other digital technologies such that they can effortlessly adopt such technologies, operate at ‘twitch speed’, are able to multitask, imagine, and visualize while communicating in multiple modalities and consequently possess higher technical skills compared to previous generations (Prensky, 2001). The term ‘digital natives’ itself assumes a generational change in digital literacies fed particularly by informal learning, opinions which are closely informed by emancipatory rhetoric surrounding the digital. As it has become used more colloquially, ‘digital natives’ suggests an inevitable increase of technical proficiency which is acquired through immersion within the digital world. Young learners are assumed to use trial and error, and an ability to network easily with peers to transfer knowledge, in the pursuit of skills which can translate to any new digital device. Implicit within this model of informal learning is the assumption that a user who develops a proficiency in one software application will be able to easily apply these skills to other forms of software.

We argue that these assumptions do not sufficiently acknowledge either the complexity of factors shaping digital literacies or the manner in which software itself ‘evolves’ in complex combinatorial patterns. For example, common affordances such as copy, cut, and paste are assumed to be naturalized and embedded across different software applications and able to be picked up with ease by this current millennial generation for work and leisure pursuits. However these affordances are poorly understood as tools contextualized within larger hierarchies of affordances and interfaces that shape both their use and more broadly people’s engagement with knowledge, culture and society in the 21st century (Livingstone, Wijnen, Papaioannou, Costa, & Grandio, 2014). Software is also poorly understood as the product of particular industrial practices, just as with other forms of technology. The more sophisticated software applications contain complex hierarchies of affordances which are regularly updated by their designers. These hierarchies in turn embody conceptual frameworks which represent a host of assumptions made by software engineers about how these applications might be deployed by users. Software development tends to have its own trajectories governed by the economic, technical and personnel dynamics of software companies. Many common forms of software, for example, demonstrate an ‘affordance creep’ over time; the multiplication of affordances and increasing complexity of interfaces that are not necessarily tied to user demand. Software, then, evolves in patterns which continually challenge users’ existing skills and understanding.

In part the term ‘digital natives’ also articulates anxieties amongst educational institutions and practitioners that they are falling behind the literacies students will bring with them to learning contexts. Educators are often encouraged to operate under the assumption that students already possess the necessary computing skills and conceptual frameworks to learn with and through generic software packages, and tend to neglect the role which software affordances themselves play in shaping how students actually ‘perform’ that software (Adams, 2006). Research indicates that such assumptions about students’ digital proficiencies are unfounded and that digital inequalities and marginalization persist around students’ access to, and use of, information and knowledge (Bennett, Maton, & Kervin, 2008; Jones & Czerniewicz, 2010; Kennedy et al., 2008). There needs to be more attention paid to distinctions between different kinds of skills and understandings within the broader digital literacies umbrella, and how these are informed by a variety of other factors in student learning. There is evidence students may not be aware of how to apply software embedded technologies effectively to enhance their learning (Khoo, Johnson, & Zahra, 2012; Valtonen, et al., 2011). In other words, digital inequality may not be defined specifically to the issue of *physical* access to digital tools (software and hardware) but may be intimately shaped by each learner’s ability to learn and apply such tools in ways which similarly constrain learning and access to knowledge, and, deeply inform the nature of any agency which an application enables. One obvious implication is that the expectation of teaching and learning institutions should perhaps be that individuals may be more or less able to exploit and critique the affordances of software and hardware (Selwyn & Facer, 2007). A student cohort where everyone has access to digital technologies does not necessarily mean that they have all naturally acquired the literacies to usefully apply, and critically engage with, those technologies. Inequalities also arise due to differences in the social and cultural support for digital use.

Labels such as ‘digital natives’ tend to conflate a basic skill in operating new technologies with broader forms of understanding and the ability to critique aspects of technology-based cultures. It is important to challenge these stereotypes and assumed competencies in order to examine the range of skills and other literacies that today’s students bring (and do not bring) to their tertiary learning. Students may not be aware of the full implications of the affordances and constraints offered through particular software; students may be as likely to be ‘captured’ by specific applications as ‘empowered’ by them. The design of software necessarily enables some practices even as it constrains and marginalizes others. An awareness

of the nature, rationale and implications of different software becomes more vital in situations where platforms and applications are assumed by their users to be ‘neutral’ tools.

Educators know very little about how students develop the skills and expertise needed to attend to the features and use of software (as application, platform, and infrastructure) in the service of completing basic and complex learning tasks. There is evidence that students adopt a range of informal approaches to meet their learning needs due to the proliferation of digital and networked tools (Peeters et al., 2014; Khoo & Cowie, 2018). Research also indicates that students’ formal software learning backgrounds are diverse (Khoo, Johnson, Torrens, & Fulton, 2011). As with other tools employed during learning, student knowledge and use of software and hardware is highly specific to their formal and informal educational, social, and cultural contexts for learning and use (Jones, Ramanau, Cross, & Healing, 2010; Valtonen, Dillon, Hacklin, & Väisänen, 2010). There is a need for further studies to raise educators’ awareness of the role that software plays in shaping and framing what and how they teach and its effects on student learning. In particular, educators can benefit from ideas and guides that help them to interrogate and reflect on their own practices - to facilitate more effective pedagogies for teaching and learning with, through and about software (including raising their students’ awareness to similar issues).

SOFTWARE LITERACY

Software studies, a comparatively new field of enquiry, proposes that ‘software’, which encompasses many forms of computer programming, is the dominant cultural technology of our time (Fuller, 2008; Johnson, 1997; Manovich, 2013). Manovich argues that we need to “think of software as *a layer that permeates all areas of contemporary societies*” (emphasis in original, 2013, p. 15), a feature of all contemporary social and cultural practices which has been comparatively neglected as an object of investigation. In an era in which our daily work routines may revolve around the use of a set of naturalized creative and communicative applications, when most social, cultural and political communication is embedded within and deeply informed by social media platforms such as Facebook, YouTube and Twitter, and a plethora of software applications run in the background of all major institutions, ‘software’ as a collective needs to become the target of greater investigation. A core premise of software studies is the need to recognize the role of software operating at a variety of levels to shape the nature of our institutions and our everyday lives. In other words, there is a vital need for software users (i.e. the vast majority in modern societies) to develop a more critical awareness of how software operates to both “empower and discipline” our social, economic, and political practices (Kitchin & Dodge, 2011, p. 10-11).

These are obviously core issues for educators but given the increasing role of software within everyday life such concerns also contain lessons for software users more widely. For instance, Adams (2007) argues for the development of more informed epistemologies of practice and advocates for software design principles that are more sensitive to pedagogical practices. And Lynch (2015) critiques the long-standing tendency to adopt new technologies without considering the deeper implications of their use to call for further investigations into the “hidden ways issues of ideology, power, and inquiry are encoded in educational software spaces” (Lynch, 2015, p. 50). Others such as van Leeuwen and O’Halloran (2011) underscore the need for software literacy modelling and use by educators and students. Being more critically aware of these aspects would allow educators and students to become more informed users of software (Vallance & Towndrow, 2007). This awareness can lead to opportunities for more equitable exploration and dialogue concerning the social, cultural and technological forces that shape the way in which knowledge within a formal discipline is developed, shared and communicated and the

implications of these aspects for teaching and learning practices. The proponents of Software Studies have tended to focus on deconstructing particular kinds of software as part of a broader effort to develop a genealogy of software as a cultural artefact - Manovich, for example, has engaged in the vital task of analyzing sophisticated media editors such as Adobe After Effects (Manovich, 2006). There remains a pressing need for empirical studies which investigate how skills and understanding around specific forms of software are acquired and applied by diverse groups of users.

In order to extend this line of research the term ‘software literacy’ is proposed and defined as the expertise in understanding, applying, problem solving and critiquing software in pursuit of particular learning and professional goals (Khoo, Hight, Cowie & Torrens, 2017). We use the term ‘literacy’ to direct attention to how individuals make use of software and for what purposes, in what contexts and with what anticipated and unexpected outcomes. Our notion of software literacy is grounded in a practice-based framework which aims to identify what distinguishes a novice user from an expert user through proposing a progressive transition from novice to expert-like understanding, capabilities and qualities (Jones, 2008; Livingstone, et al., 2014). Our framework clarifies three progressive tiers of development towards software literacy. They are:

1. A basic skill level where a learner can use a particular software. This typically means using common affordances provided by default at the top level of a hierarchy of affordances (for example, cut, copy and paste functions which tend to be featured in prominent positions within many interfaces for creative and communication applications),
2. An ability to independently troubleshoot and problem-solve issues faced when using the full range of affordances constructed by the software. When teaching software this is a key threshold reached by students; the point where they recognize their own novice level of proficiency in using an application but are able to identify and make use of resources which can allow them to understand and apply further sets of affordances. For more sophisticated forms of software, users may never achieve a level of ‘mastery’ over an application or system (they may not know of, understand and can apply each and every one of its affordances). And they may find their existing skills and understanding continually challenged by periodic updates to its hierarchies of affordances and/or interface. But these users nevertheless have acquired the ability to problem-solve how to apply that software to new tasks.
3. And finally, the ability to critique the software, including being able to apply such critique to a range of applications designed for a similar purpose and to further use these understandings for new software learning. In immediate terms, users at this level have an understanding of the conceptual framework underlying an application; i.e. the archetypal user it is aimed at, and the kinds of practices it is primarily designed for, and hence what that application might or might not be most usefully directed towards. Users at this level, then, develop a knowledge and an understanding of a range of software and what is ‘useful’ for particular kinds of tasks.

Implicit within this tiered framework is also a graduated sense of how users’ agency is enabled and constrained through their use of software in different contexts. At the first tier, users are more likely to adapt their existing practice to suit what they are able to achieve using an application. Successfully completing a specific set of tasks might be all that these users require and they might see no need to explore that software further. In a sense, these users might be seen to have been ‘captured’ by that particular application; they are yet to develop a richer awareness of the possibilities to deploy that software

for other purposes, nor have an understanding of how competing applications may offer different possibilities for their existing practices.

Users at the second tier can be assumed to have developed greater agency in relation to an application or platform. Enabled by their ability to troubleshoot when they encounter difficulties in making the application do what they want it to, they are perhaps more likely to feel empowered to explore the wider possibilities of that application. Users at this level appear to also be more likely to start using a software application or platform in ways that would be surprising to its original programmers. Moving beyond the assumptions about core users held by that applications' designers, they are able to use an application in ways that move beyond the default practices most easily enabled within its hierarchy of affordances.

Finally, users at the third tier of software literacy are assumed to be able to explore their agency across a range of software applications or platforms. As more critical users of software, they are more likely to develop an awareness of the range and variety of practices which are enabled and constrained by particular sets of software. They potentially have the ability to understand what an application means in detail for their particular practices; that is, what each application offers them, what each application 'does well' and what it does poorly. They have agency to pick and choose the software tools that are most appropriate for a specific practice and to apply sets of applications together toward particular ends. They understand the imperfections of any given piece of software but also the ways in which these might be used for their own purposes. For those users who develop a 'mastery' in a particular sophisticated set of applications or platform, they are capable of generating a critique of that software as it evolves over time (over a series of upgrades). Where a user at the first tier of software literacy might not be able to easily move from one application once they achieve a basic proficiency in its use, users at this third tier are much more easily able to 'abandon' software which is no longer useful and to commit to learning a new application or platform, and to more easily apply acquired skills and understanding toward this end.

A key point to emphasize here is that these three tiers are intended as reference points for thinking through software literacies. There is inevitably a fluidity and flexibility across these tiers for any group of users. Any single user, for example, might operate across more than one level, having gained basic proficiency in one application or applications but not necessarily having the ability to immediately transfer this skill set to another kind of application or platform. Complex applications tend to develop their own hierarchies of affordances, and are not necessarily easily understood by new users, even those with some skill in other forms of software. For example, gaining a mastery of the Adobe Creative Suite does not mean that users have therefore also achieved a mastery of other complex software applications such as Computer-Aided Design systems.

This framework also assumes a variety of skills and understandings are constituted in what it means to be fully 'literate' in software. A key threshold within tier 3 is the point where users are able to transfer a critical understanding of software affordances and use to new software and to familiar software in new contexts and situations. This knowledge might be relatively contained to a specific set of applications that a user encounters most often - such as students acquiring a rich and nuanced awareness of the kinds of applications which, inform, enable and can extend their practice within a particular field. Many but perhaps not all of these users can be assumed to then have the sophisticated understanding needed to consider how software (at various levels of application, platform and infrastructure) more broadly shapes, enables and constrains the particular kinds of knowledge and actions that are readily available to them. At this level, students (and by extension, citizens) are assumed to be capable of recognizing more broadly the role which 'software' plays in various levels within contemporary life, and to more deliberately choose platforms and applications based on a thorough understanding of the conceptual and practical implications of their choices. As with other forms of digital literacies, this third tier accentuates

an essential ability to ‘critique’ software as it appears more broadly across other everyday practices. This is an attribute also recognized by other researchers as a key characterization of 21st century learners and working professionals to participate productively in the political, economic, leisure and social spheres of everyday life (Gilbert 2005, Jenkins et al. 2009, UNESCO, 2015).

Developing one’s software literacy relies on a combination of general competency with software and technologies, together with the ability to undertake more independent (especially informal) learning of discipline-specific knowledge as and when required. This conceptual framework is a response to current digital literacy frameworks and related terms which do not go far enough to identify the implications of the choice of software applications and platforms on what can be achieved (Hegarty et al., 2010; Livingstone et al., 2013). The software literacy framework outlined here is intended to complement existing research on digital literacies; to more properly acknowledge the emergence of software as a cultural artefact possessing agency within social, cultural, economic and political practices (Fuller, 2008; Johnson, 1997; Manovich, 2013).

SOLUTIONS AND RECOMMENDATIONS

The potential of our three-tier software literacy framework was evidenced through a case study of university context teaching and learning environment. The aim was to explore, examine, and theorize on how the notion of software literacy is understood, developed, and applied in university teaching and learning contexts, and the extent to which this understanding is useful when translated into new contexts of learning with and through software (Khoo, Hight, Torrens, & Cowie, 2016). To address this broader aim, the study, confined to the disciplines of media studies and engineering, sought to explore the nature of informal and formal learning of software applications as an entry point in investigating the nature of software literacies fostered within distinct tertiary student cohorts. The study initially focused on student learning of PowerPoint as a generic form of software application, typically learned partially or wholly through informal learning, before progressing to examining student learning of disciplinary-specific software learning in media studies (e.g. Adobe Photoshop, Final Cut Pro), and engineering (e.g. SolidWorks, a 3D computer-aided design (CAD) application). Within both these disciplines, PowerPoint was a common feature of lecture sessions, but not specifically examined as software. While engineering students were required to learn SolidWorks as part of disciplinary knowledge, the media studies students encountered software as part of production electives within a broader degree programme and were encouraged to use a range of applications as needed for creative projects. Each of these disciplines required students to apply professional-level creative and editing applications toward learning tasks, and provided detailed instruction in software which was both comparatively complex and relatively unknown to incoming student cohorts. Each disciplinary context also provided distinct sets of knowledge about software provided to students, contrasting experience with a range of alternative applications to those taught in workshops, but nevertheless shared expectations of the need for students to employ informal learning to complete their understanding of specific pieces of software. The overall objective was to understand the impact of more direct instruction (of sophisticated software applications) on student software literacy development and disciplinary understanding.

The findings of this study support the existence of the three-tier software literacy framework but indicate there is some flexibility in student development and movement between the tiers. It was demonstrated that student ability to achieve the higher tiers of software literacy did not preclude them from needing to revisit earlier levels in contexts where they encounter new but similar software. As such,

developing a relatively sophisticated and critical understanding of an application or platform does not necessarily transfer directly to the learning of other forms of software. This suggests educators should not assume student competency with software across contexts (e.g., informal to formal, from campus to workplace settings). The framework provides a key conceptual tool for practitioners in terms of understanding the role of troubleshooting, for example, as an important development stage in learning with and through software.

There was evidence of students having arrived at tier 3 (critical literacy) as well as those not moving beyond tier 2 (troubleshooting). For example, student (mis)assumption that the course lecture notes presented in PowerPoint's bullet point format in and of themselves constitute an adequate elaboration of pertinent disciplinary knowledge. The affordances and constraints that PowerPoint and discipline-based software offer in shaping disciplinary knowledge needs further consideration by students and by educators. Educators therefore need to be aware of the implications of their choice and modelling of a software application or platform. Findings suggest they would be wise to introduce and help students to develop critical awareness of how such software can inform and shape their own understanding of disciplinary knowledge and practice.

These findings were manifested in distinct but closely aligned patterns within each disciplinary context. However, there was also a difference in the way each discipline prioritized software literacy development as part of their student learning pathway and graduate profiles. For example, in the media studies case, critical thinking as emphasized in the third tier is seen as a core aspect of creative disciplinary knowledge. Being able to judge the creative capacities of competing software has implications for the nature of the practice which is being developed, the form and eventual conceptual complexity of media products themselves, as well as more practical considerations such as budget. In contrast, in the engineering case, students are expected to develop digital and software literacy skills as part of their professional learning. For graduates, the quality of engineering design produced and a design's efficiency and effectiveness are prioritized over the kinds of creative reflexivity observed in the media studies case study. Such disciplinary differences reveal a need for each discipline to continually reflect upon how discipline-specific software teaching and learning is positioned in relation to graduate profiles and the breadth and depth of literacies required to engage successfully with software.

Another clear finding reiterates that multiple learning pathways exist for exploring the affordances of any particular software, both formal and informal. Students tend to prefer informal strategies as a supplementary to and at times above formal strategies to learn discipline-based software (these included YouTube video tutorials, along with peer networks) (Khoo & Cowie, 2018). Educators could usefully be informed by and take advantage of students' informal repertoire of learning strategies and networks including their accessing of (web-based) resources and pooling of peer expertise. Educators drawing from students' already established informal learning strategies would recognize the relevance of the social and cultural context in shaping effective technology and software engagement.

The findings importantly challenge current notions and assumptions of the digitally literate generation. When learning to use PowerPoint most students perceived themselves to be competent and confident early adopters of technologies and could identify the general affordances of generic software (PowerPoint) to the extent of critiquing the design and performance of PowerPoint practice in their lectures. They were able to recognize discipline-specific affordances among the applications formally taught and acknowledged these to be central in their engagement with disciplinary knowledge. However, very few were able to critique how the design of the 'software' itself might shape their disciplinary knowledge. Very few students demonstrated critical thinking about the nature and role of the software they were using and most were not able to describe or discuss applications beyond those used in their learning. That is, there

was very little evidence of tier 3 software literacy. Students' superficial critique of software challenges current assumptions that today's students as 'digital natives' have developed critical awareness simply through familiarity with and regular use of a variety of everyday applications and platforms.

The diversity of student cohorts and the range of understandings and the variation in familiarity, skills and experience students bring with them to the formal software learning context constitute a further challenge for teaching of, through and about software. Some students may already have a critical orientation towards software (or, more frequently, toward specific applications). And there is demonstrably an advantage in terms of more advanced software learning for those who have prior experience with other software with a similar conceptual framework. In response to this diversity educators could usefully direct time and attention to formatively assessing students' initial software literacy and adapting teaching activities in light of this. Most crucially, however, the overall findings indicate there is no single best approach (one size fits all) to teaching discipline-specific software. Educators adopting a range of teaching approaches (formal and informal) and being flexible to address diverse learning needs represents a crucial part of supporting student learning.

FUTURE RESEARCH DIRECTIONS

Software is a site of cultural production and an actor in social, economic, political practices; it needs to be analyzed in terms of its embedded values and effects on such practices. The brief summary of case study findings outlined above is limited due to the situated nature of the investigation into discipline-specific software teaching and learning. The participants in the study represent a convenience purposive sample of educators and students from one educational institutional setting. They were from two distinctly different disciplinary contexts, with distinct and different disciplinary foci and expectations of software teaching and learning. Educators were also careful to point out the study only focused on some courses within a program and all universities have different interpretations of how software teaching and learning can be enacted. Many educational practices are now intimately associated with software-based tools and applications, suggesting an urgent need for research which explores the implications for such practices and the manner in which they operate and are embedded within specific institutional, disciplinary and learning contexts.

CONCLUSION

Software literacy is an essential part of learning and living in the 21st century; something which transcends the use of any particular tool and any particular educational, social and cultural context. Software is an increasingly central part of the palette of understandings and skills which comprise the broadening umbrella of digital literacies. The software literacy framework outlined here offers a basis for addressing the social and cultural dimensions of software-centered learning, including the interactions between formal tertiary environments and informal learning networks.

The research outlined above addressed a tertiary educational context and a focus on specific sets of applications already embedded within existing industrial practices (from creative workflows suggested by media production practices, through to the 3D CAD practices that are widely used across design, architecture and engineering industries). Outside of this formal learning environment, software offers a challenging and rapidly evolving part of contemporary life. Everyday applications used for word

processing and other creative and communication practices generate a low-level demand to maintain skill proficiency in the face of continually updated software packages. Software interfaces are proliferating across a variety of other aspects of everyday spaces, from control panels for services in the urban landscape, to their increasing ubiquity across household appliances and a host of everyday encounters within retail, transport and entertainment services. These also have implications for practices that are at the core of and related to how knowledge is generated and curated. Contemporary museum and library environments, for example, are increasingly constituted through software platforms - in ways which are not being researched as rapidly as they are being pursued.

Other examples of how software tools can influence the way we live are not hard to find. There is an increasing reliance locally and internationally on a host of practices and specializations within the field of data science (such as the use of arrays of sensors to generate data fields, data mining and the new frontier of deep learning). These are being used within policing, environmental, agricultural sectors for making sense of data in ways which can inform and drive policy and funding decisions. Newer media platforms such as augmented reality (AR) and virtual reality (VR) (collectively referred to as XR) have yet to become as fully embedded within education, entertainment and knowledge industries as they are within military-industrial contexts, but AR is increasingly a default platform for training practices and on-site support materials across a huge range of applied practices (from mechanical to medical). As of this writing there are a host of experimental applications which provide cultural encounters for users which are fully encased within dynamic, fluid and rapidly evolving software environments within the XR spectrum. More seriously, the threshold for access by everyday users to increasingly sophisticated creative production platforms is becoming lower, at the same time as the more disturbing implications for some of these platforms are becoming clearer. There is already unease, for example, around the potential of 'deepfakes', a form of audio-visual production aligned to machine learning techniques which allows the easy generation of fabricated video content. Public figures particularly within the United States have been early targets but commentators are speculating on the implications for all of us if these forms of 'synthetic' media are added to an existing online climate of misinformation and disinformation (Marwick & Lewis, 2017; Wardle & Derakhshan, 2017).

There is obviously a pressing need for all of us as citizens to have a critical understanding of software to allow us to make more informed choices about their use, to enable us to transfer this critical understanding to software we have yet to encounter, and most crucially to understand that all software has nuanced affordances and limitations. This has implications for education providers, especially universities, in their role in fostering critical thinking and serving as a key critic and conscience of society. It is crucial to ensure all students and educators are supported in teaching and learning processes when these are mediated through and focused on software. As society moves to exploit the potential of a range of 'coded' practices there is a need for further detailed empirical investigation into the extent to which learners of applications, platforms and infrastructures become software literate. In formal and informal settings this is needed to ensure equitable and critical learning with, through and about software.

ACKNOWLEDGMENT

The authors gratefully acknowledge funding support from the Teaching and Learning Research Initiative, New Zealand Council for Educational Research, Wellington, New Zealand.

REFERENCES

- Adams, C. (2006). PowerPoint, habits of mind, and classroom culture. *Journal of Curriculum Studies*, 38(4), 389–411. doi:10.1080/00220270600579141
- Adams, C. (2007). On the “informed use” of PowerPoint: Rejoining Vallance and Towndrow. *Journal of Curriculum Studies*, 39(2), 229–233. doi:10.1080/00220270601175246
- Ala-Mutka, K. (2011). *Mapping digital competence: Towards a conceptual understanding*. Seville: JRC-IPTS. Retrieved from <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=4699>
- Alexander, B., Adams Becker, S., Cummins, M., & Hall Giesinger, C. (2017). *Digital literacy in higher education, Part II: An NMC horizon project strategic brief (Volume 3.4)*. Austin, TX: The New Media Consortium. Retrieved from <https://blog.stcloudstate.edu/ims/files/2017/08/2017-nmc-strategic-brief-digital-literacy-in-higher-education-II-ycykt3.pdf>
- Beetham, H., & Sharpe, R. (2010). *Digital literacy framework*. Retrieved from <http://jiscdesignstudio.pbworks.com/w/page/46740204/Digital%20literacy%20framework>
- Bennett, S., Maton, K., & Kervin, L. (2008). The ‘digital natives’ debate: A critical review of the evidence. *British Journal of Educational Technology*, 39(5), 775–786. doi:10.1111/j.1467-8535.2007.00793.x
- Bulfin, S., & Koutsogiannis, D. (2012). New literacies as multiply placed practices: Expanding perspectives on young people’s literacies across home and school. *Language and Education*, 26(4), 331–346. doi:10.1080/09500782.2012.691515
- Fuller, M. (2008). *Software studies: A lexicon*. The MIT Press. doi:10.7551/mitpress/9780262062749.001.0001
- Furlong, J., & Davies, C. (2012). Young people, new technologies and learning at home: Taking context seriously. *Oxford Review of Education*, 38(1), 45–62. doi:10.1080/03054985.2011.577944
- Gilbert, J. (2005). *Catching the knowledge wave? The knowledge society and the future of education*. NZCER Press.
- Goodfellow, R., & Lea, M. R. (2014). *Literacy in the Digital University: Critical perspectives on learning, scholarship and technology*. Routledge.
- Hague, C., & Logan, A. (2009). *A review of the current landscape of adult informal learning using digital technologies*. Retrieved from http://preview.futurelab.org.uk/resources/documents/project_reports/becta/Adult_Informal_Learning_educators_report.pdf
- Hegarty, B., Penman, M., Kelly, O., Jeffrey, L., Coburn, D., & McDonald, J. (2010). *Digital Information Literacy: Supported Development of Capability in Tertiary Environments*. Wellington, New Zealand: Ministry of Education. Retrieved March 10, 2012 from http://www.educationcounts.govt.nz/publications/tertiary_education/80624
- Jenkins, H., Purushotma, R., Weige, M., Clinton, K., & Robison, A. J. (2009). *Confronting the Challenges of Participatory Culture - Media Education for the 21st Century*. Cambridge MA: MIT Press. Retrieved June 20, 2012 from http://mitpress.mit.edu/sites/default/files/titles/free_download/9780262513623_Confronting_the_Challenges.pdf

JISC. (2014). *Developing digital literacies*. Retrieved <https://www.jisc.ac.uk/full-guide/developing-digital-literacies>

Johnson, S. (1997). *Interface Culture: How New Technology Transforms the Way We Create and Communicate*. HarperCollins Publishers.

Jones, C., & Czerniewicz, L. (2010). Describing or debunking? The net generation and digital natives. *Journal of Computer Assisted Learning*, 26(5), 317–320. doi:10.1111/j.1365-2729.2010.00379.x

Jones, C. M. (2008). From novice to expert: Issues of concern in the training of psychologists. *Australian Psychologist*, 43(1), 38–54. doi:10.1080/00050060601089470

Kennedy, G., Judd, T. S., Churchward, A., Gray, K., & Krause, K.-L. (2008). First year students' experiences with technology: Are they really digital natives? *Australian Journal of Educational Technology*, 24(1), 108–122.

Khoo, E., & Cowie, B. (2018). Trial-and-error, Googling and talk: Engineering students taking initiative out of class. In D. Corrigan, C. Bunting, A. Jones, & R. Gunstone (Eds.), *Navigating the changing landscape of formal and informal science learning opportunities* (pp. 193–206). Springer. doi:10.1007/978-3-319-89761-5_12

Khoo, E., Hight, C., Torrens, R., & Cowie, B. (2016). *Copy, cut and paste: How does this shape what we know? Final report*. Wellington, New Zealand: Teaching and Learning Research Initiative. Retrieved from <http://www.tlri.org.nz/tlri-research/research-completed/post-school-sector/copycut->

Khoo, E., Hight, C., Torrens, R., & Cowie, B. (2017). *Software Literacy Education and Beyond*. Springer. doi:10.1007/978-981-10-7059-4

Khoo, E., Johnson, M., & Zahra, A. (2012). I learnt a whole lot more than churning out an essay: Using online tools to support critical collaborative inquiry in a blended learning environment. *Journal of Open, Flexible and Distance Learning*, 16(1), 127–140.

Kitchin, R., & Dodge, M. (2011). *Code/Space: Software and Everyday Life*. The MIT Press. doi:10.7551/mitpress/9780262042482.001.0001

Kvavik, R. B. (2005). Convenience, communications, and control: How students use technology. In D. Oblinger & J. Oblinger (Eds.), *Educating the Net generation* (pp. 7.1–7.20). Retrieved from <http://www.educause.edu/research-and-publications/books/educating-net-generation/convenience-communications-and-control-how-students-use-technology>

Livingstone, S., Wijnen, C., Papaioannou, T., Costa, C., & Grandio, M. (2014). Situating media literacy in the changing media environment: critical insights from European research on audiences. In N. Carpentier, K. Christian Schröder, & L. Hallet (Eds.), *Audience transformations: Shifting audience positions in late modernity* (pp. 210–227). Routledge.

Lynch, T. R. (2015). *The Hidden Role of Software in Educational Research: Policy to Practice*. Routledge Research in Education. Routledge. doi:10.4324/9781315751177

Manovich, L. (2006). After Effects or the Velvet Revolution. *Millennium Film Journal*, 45/46, 5–19.

Manovich, L. (2013). *Software takes command*. *International Texts in Critical Media Aesthetics*. Bloomsbury Academic Plc.

- Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. New York: Data & Society Research Institute. Available from <https://datasociety.net/output/media-manipulation-and-disinfo-online/>
- Oblinger, D. (2003). Boomers, gen-Xers, and millennials: Understanding the “new students. *EDUCAUSE Review*, 38(4), 36–45.
- Peeters, J., Backer, F. D., Buffel, T., Kindekens, A., Struyven, K., Zhu, C., & Lombaerts, K. (2014). Adult learners’ informal learning experiences in formal education setting. *Journal of Adult Development*, 21(3), 181–192. doi:10.1007/10804-014-9190-1
- Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5), 1–6. doi:10.1108/10748120110424816
- Sefton-Green, J., Nixon, H., & Erstad, O. (2009). Reviewing approaches and perspectives on ‘Digital Literacy’. *Pedagogies*, 4(2), 107–125. doi:10.1080/15544800902741556
- Selwyn, N., & Facer, K. (2007). *Beyond the digital divide: Rethinking digital inclusion for the 21st century*. Futurelab. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.3384&rep=rep1&type=pdf>
- Tapscott, D. (1999). *Growing up digital: The rise of the net generation*. McGraw-Hill.
- Tapscott, D. (2009). *Grown up digital*. McGraw-Hill.
- Underwood, J. (2009). *The impact of digital technology*. BECTA. Retrieved from <http://publications.becta.org.uk/display.cfm?resID=41343>
- UNESCO. (2015). *Leveraging information and communication technologies to achieve the Post-2015 Education goal: report of the International Conference on ICT and Post-2015 Education*. Paris: UNESCO. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000243076>
- United Nations Educational, Scientific, and Cultural Organization (UNESCO). (2013). *Global media and information literacy (MIL) assessment framework: Country readiness and competencies*. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000224655>
- Vallance, M., & Towndrow, P. A. (2007). Towards the ‘informed use’ of information and communication technology in education: A response to Adams’ ‘PowerPoint, habits of mind, and classroom culture’. *Journal of Curriculum Studies*, 39(2), 219–227. doi:10.1080/00220270601105631
- Valtonen, T., Pontinen, S., Kukkonen, J., Dillon, P., Väisänen, P., & Hacklin, S. (2011). Confronting the technological pedagogical knowledge of Finnish net generation student teachers. *Technology, Pedagogy and Education*, 20(1), 1–16. doi:10.1080/1475939X.2010.534867
- van Leeuwen, T., & O’Halloran, K. L. (2011). *Towards a social theory of semiotic technology: Exploring PowerPoint’s design and its use in higher education and corporate settings*. Retrieved from <http://datasearch2.uts.edu.au/cpce/projects/detail.cfm?ProjectId=2008000350>
- Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe report, DGI (2017)9. Available from <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>

ADDITIONAL READING

Edwards, E. (2015). Software and the hidden curriculum in digital education. *Pedagogy, Culture & Society*, 23(2), 265–279. doi:10.1080/14681366.2014.977809

van Dijck, J. (2013). *The Culture of Connectivity: A critical history of social media*. Oxford University Press. doi:10.1093/acprof:oso/9780199970773.001.0001

KEY TERMS AND DEFINITIONS

Affordance: A person’s perceived opportunity to utilise a particular tool for action, for example, a doorknob is for turning. Within software, these are typically organised through user interfaces.

Digital Literacy: An umbrella term for a number of competencies, skills and understandings associated with digital infrastructural technologies (we argue software literacy falls within this same umbrella).

Pedagogy: Study and practice of teaching.

Software: Machine readable instructions which directs a computer’s processor to perform specific operations. For everyday software applications, these instructions are put into operation as users run an application on any digital, networked, or mobile device.

Software Application: A computer program designed to perform a set of tasks or functions. Common everyday examples range from an ‘app’ on a smartphone, to desktop computer programs such as a web browser.

Software Literacy: A person’s expertise in understanding, applying, problem solving and critiquing software in pursuit of particular learning and professional goals.

Software Platform: A coherent programming environment or system which supports applications. Examples include operating systems such as iOS or Android, or social media platforms such as Facebook.

ICT Investments on Economic Sectors With International Comparative Advantage and the Diffusion of Prosperity

Ioannis Papadopoulos

Metropolitan College, Greece

Apostolos Syropoulos

 <https://orcid.org/0000-0002-9625-1482>

Greek Molecular Computing Group, Greece

INTRODUCTION

It is well documented in the scientific literature that countries facing economic crises, or the years following that difficult period, need extensive investments (domestic or foreign) in various economic sectors in order to re-gain their international competitiveness and to trigger long-term and viable economic development. Such investments are critical factors to increase nation's GDP, further reduce unemployment rate, increase purchasing power and spread the benefits of economic prosperity across all levels of society. A crucial aspect is to examine to what extent information and computer technologies (ICTs), as they incorporated in this investment process, can accelerate and leverage the positive impact in the economic circuits and provide an added value in the economy. Certainly, such investments would be highly effective to take place in economic sectors that appear international comparative advantage in conjunction with relative interventions in public sector.

It has been long investigated and empirically proved worldwide that adoption of information technologies from private and public sector can provide immediate, direct and indirect benefits in short and medium-long term basis. As a demonstration of this we will examine the case of Greece, a country that faced an unprecedented economic crisis for the western world standards, and we will explain how adoption of information technologies can be realized in both the private and the public sector, so as having a positive multiplying effect in economy with positive results for the whole society. Through this analysis, we shall underline the importance of ICT investments on economic sectors with comparative advantage in Greece that potentially can accelerate the current economic development trend and provide solutions that will create income and opportunities for all, especially for the new generation.

BACKGROUND: ECONOMIC CRISIS IN GREECE (2009-2017)

Before analyzing the main topic of this article, it is fundamental to briefly mention the background of Greece in terms of the economic crisis. Greece has been chosen as a unique example of a western world country that has been suffered an economic crisis of such magnitude. Economic crisis triggers negative social phenomena across all levels of economic function. Indicatively, the country under crisis suffers reduced GDP (Gross Domestic Product - the total amount of values produced in economy in a year)

DOI: 10.4018/978-1-7998-3479-3.ch114

that is usually translated into potential personnel reductions (or in the best case there are no new hirings), decreased salaries and pensions and reduced budget of public investments (new roads, hospitals, schools, etc.), amongst others. Accordingly, in the private sector, firms face reduced demand for their products, difficulties in borrowing money from banks, increased tax payments and other negative facts that lead them to reduce personnel, reduce their program of investments and holding money because of the uncertainty prevailing in economy.

The figures that follow clearly depict the impact of the crisis on the fundamental economic figures of Greece. These figures indicate the economic tunnel that Greece has been through for almost a decade with major implications in society and in economy.

Figure 1 shows the Greek GDP over the last 40 years. The Greek economy, achieved a historical economic progress from 1974 onwards, which violently halted in 2009 signaling the collapse of the economic process. Greece, from 2009 until 2016, lost almost 25% of its income.

Figure 1. The evolution of the Greek GDP over the last 40. Source: World Bank, Gross Domestic Product for Greece [MKTGDPGRA646NWDB], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/MKTGDPGRA646NWDB>, June 13, 2019.

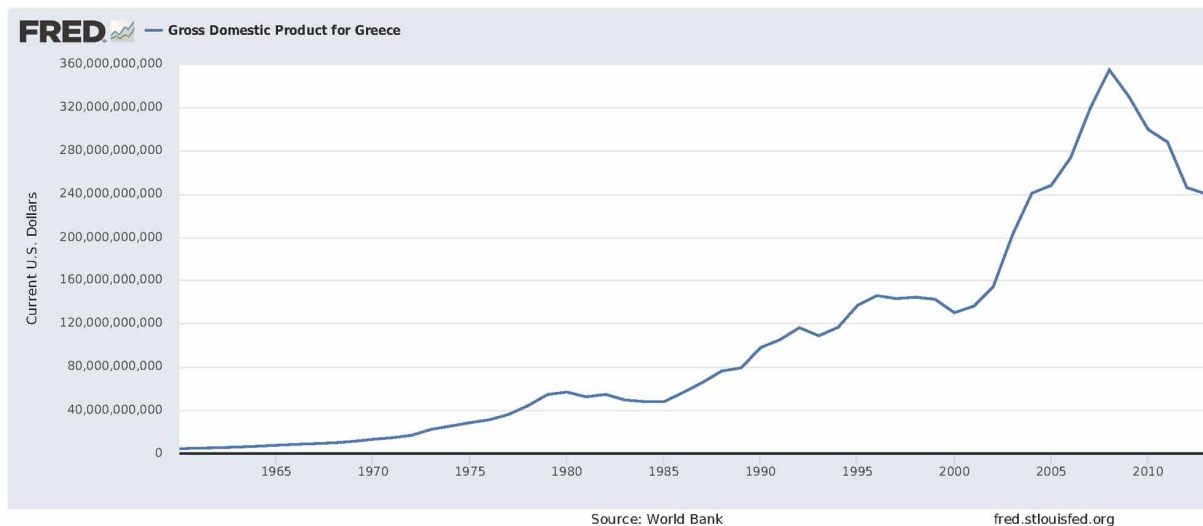


Figure 2 shows the same negative situation in terms of income per capita, indicating the worsening of the life-conditions for the Greek citizens.

During the same period, one of the most negative implications in such a difficult economic environment was the sky-rocketing of unemployment rate, which touched the level of 28% in 2013. This is clearly shown in figure 3.

The previous figures illustrate the unprecedented magnitude of the economic crisis in Greece and its dramatic impact on society. Those figures show the recessionary background of the Greek economy and the same time the starting point for the future challenges and prospects lie ahead and that will be analyzed over the next chapters. The dimensions of the crisis in Greek economy are not in the scope of this article [the interested reader should consult (Papadopoulos & Syropoulos, 2018) for more details].

Figure 2. Evolution of GDP per Capita for Greece. Source: World Bank, Gross Domestic Product Per Capita for Greece [PCAGDPGRA646NWDB], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/PCAGDPGRA646NWDB>, June 13, 2019.

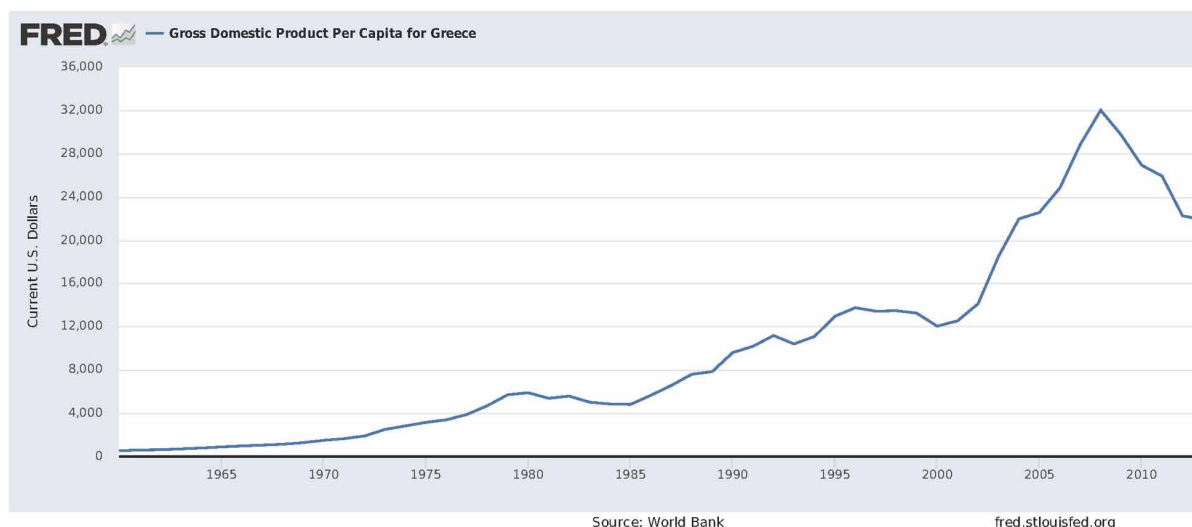


Figure 3. Evolution of unemployment rate for Greece. Source: Organization for Economic Co-operation and Development, Unemployment Rate: Aged 15-64: All Persons for Greece [LRUN64TTGRQ156S], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/LRUN64TTGRQ156S>, June 13, 2019.



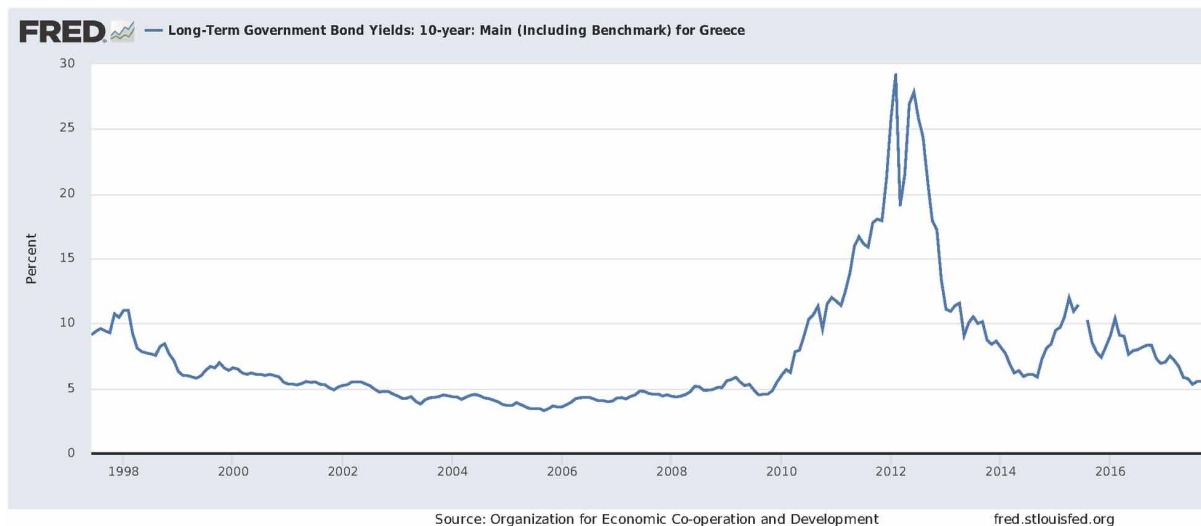
ECONOMIC MOMENTUM

The opposite of crisis is economic development. The term “development” means the expansion of an economy and the increasing of the level of wealth and the level of well-being in society through new well-paid jobs and new public and private investments, amongst others. Especially for Greece, after a long period of 10 years in crisis, economic development is a one-way road for the future of the country.

Many opinions have been expressed in public dialogue and in scientific literature about the necessity for a strong and viable development to be triggered in the Greek economy after so many years of recession and dramatic social implications. But how this economic development can be realized? Certainly, the key-word here is the term investment. Development of an open economy can only be achieved through investments which accelerate and multiply the economic growth. In simply words, an investment is a dedicated amount of money used to create something, with the hope that this will return in the future a profit supporting the economy in its expansion. Investments are separated in public (roads, schools, energy infrastructure, hospitals, etc.) and in private ones (new buildings, new businesses, new production lines, collaboration with public sector, etc.).

Economic rationalism and most modern economic theories underline the fact that investments should be realized in sectors that appear historical and/or potential comparative advantage at a regional or international level. In this way, investments could achieve a high efficiency and a significant profitability rate leading the economy to higher levels of balance with better opportunities for all citizens to improve their lives. The only question here is the following one: Is Greece ready to proceed to that kind of investments through a nationally accepted strategic development plan? Figure 4 shows the evolution of the 10year Greek government bond yield. After the tremendous levels of 2010-2013 and 2015, a constant downward trend has been initiated from 2016 onwards indicating a relative stability in the Greek economy and the regaining of the trust from the global investors about the prospects of the Greek economy.

Figure 4. Evolution of the 10y Greek government bond yield. Source: Organization for Economic Co-operation and Development, Long-Term Government Bond Yields: 10-year: Main (Including Benchmark) for Greece [IRLTLT01GRM156N], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/IRLTLT01GRM156N>, June 14, 2019.



In addition, figure 5 shows the general trend of the exports in Greece (goods & services) from 1996 onwards. The most impressive and promising for the future dimension here is that Greek exports, after the crisis shock in 2009, are constantly increasing their magnitude in terms of value. This indicates the extroversion of many Greek companies even in the middle of the crisis and the adjustment of those firms in this new and challenging environment.

Figure 5. Evolution of exports of goods and services in Greece. Source: Organization for Economic Co-operation and Development, Exports of Goods and Services in Greece [GRCEXPORTADSMEI], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/GRCEXPORTADSMEI>, June 13, 2019.



Finally, the Greek fleet (dead-weight tonnage), which is the biggest in the world [see (United Nations, 2018)], and tourism are two economic sectors with significant comparative advantage at international level. In particular, Greek tourism (see figure 6) shows constant increase in arrivals and receipts over the last years.

All previously mentioned critical elements strongly indicate that for Greece now is the right momentum to come back to a strong economic development after so many years in economic crisis. Additionally, one might say, that this development can now be conducted under better terms than the development that was taken place before 2009 because of the massive reforms that have been made in public sector along with relative legislative production.

The only crucial issue here is that Greek politicians must realize and set to implementation a holistic and strategic, national scale development plan. One that will use the right resources and ICT investments at sectors with international comparative advantage that eventually will trigger strong economic development. This development must be guided at a direction that will reinforce social cohesion, reduce inequalities while in the same time will create high quality well-paid jobs with prospects.

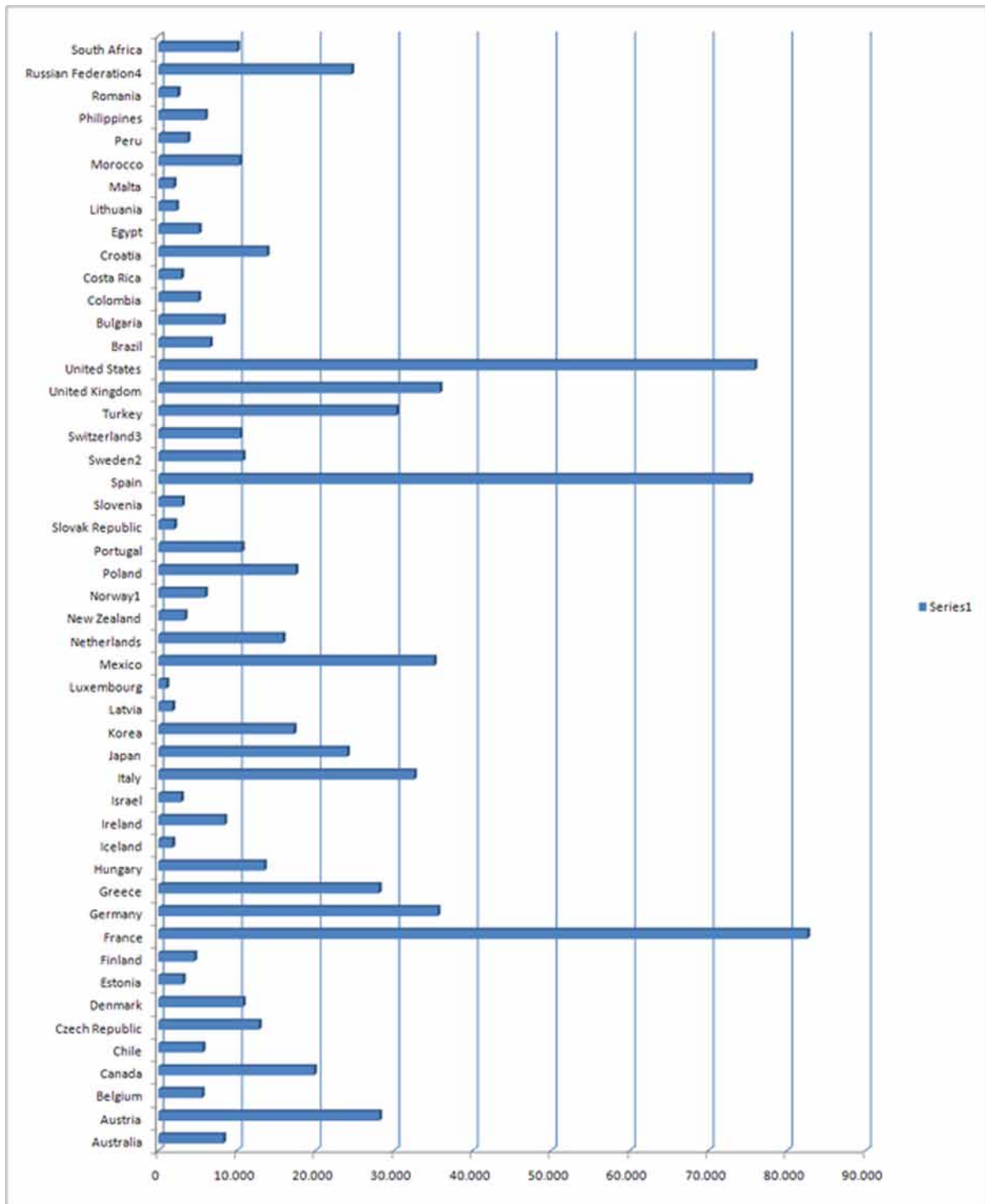
Over the next sections, we will focus exactly on this kind of analysis, meaning that a viable and strong economic development can be triggered by using extensive ICT investments in sectors with comparative advantage in Greek economy.

HOW TO ACHIEVE ICT COMPETENCE

In order to produce personnel, scientists, and engineers who have exceptional ICT skills, it is necessary to design a national educational strategy that will provide the required knowledge. This means that all primary education pupils will have to learn a number of topics in order to be able to deepen their knowledge later in middle school, high-school, and university. Currently, Greece does not have a curriculum

Figure 6. International tourist arrivals, 2012-16. Source: OECD Tourism Trends and Policies 2018 (data from <https://dx.doi.org/10.1787/888933639474>).

13



that aims to make pupils well versed in ICT. Instead, ICT lessons are considered something where pupils take it easy. In addition, the grading system makes it impossible to give fair grades and so only very few pupils who are really interested in ICT do participate in any kind of school actions [this a more general problem that is described in full detail in (Côté & Allahar, 2007)]. The worst part is that the university admission system is designed in such a way that students are admitted to university departments that have nothing to do with their skills and their background. Naturally, this has to change.

The modern trend is to use Open Source software so as to reduce the cost of computer systems. In addition, this will create an ecosystem where people will work to improve software and/or create new software to solve existing or new problems. Thus, we favor an educational approach that uses a set of Open Source programs and tools as learning vehicles. Naturally, this includes Open Source operating systems (e.g., Linux, OpenBSD, OpenIndiana, etc.). Thus, pupils of elementary education should be able to prepare documents, spreadsheets, and presentations. Depending on the available resources, one can use little robots or anything else (i.e., Kojo) to introduce pupils to LOGO style programming. In middle-school, pupils should be introduced to programming. We think that Python is a good tool to introduce pupils to programming. In three years time, one can start from very simple topics and go on with GUI programming which is both more demanding and rewarding. At high-school, pupils who aim to study information sciences and related fields, should get a deeper understanding of a number of topics [e.g. script programming with Scala (Loverdos & Syropoulos, New York)], while all pupils could be introduced to HTML and related technologies. Even learning how to use LaTeX [e.g., see (Syropoulos, Tsolomitis, & Sofroniou, 2003)] would be quite beneficial for pupils.

COMPARATIVE ADVANTAGES, ICT ENHANCEMENT & ECONOMIC DEVELOPEMENT

ICT investments on sectors with international comparative advantage can firstly triggering and later on accelerating economic development. This is so, because based on economic theory approaches, comparative advantage means that a country can produce added-value easily since the production factors (Labor, Land, Capital) are cheaper to acquire and use in the production process. This is the core of a viable, long-term economic development since the Greek economy can exploit a double nature: sectors with international economic comparative advantage with the assistance of ICT investments. ICT investments in such sectors can certainly facilitate and accelerate the economic process. Such technologies could potentially include e-tools for the support of exports companies, e-tools to improve agricultural methods and productivity, online banking tools interconnected with public authorities and finally tertiary level education e-platforms and e-tools for remote education solutions. Additionally, we can add in this list massive investments on renewable energy networks that are centrally and electronically supervised at a national level so as to maximize productivity and allocation of benefits, or new generation e-platforms for tourism to promote attractive and under-advertised destinations-disruptors along with real-time reservation abilities. As for public sector this approach could include total interconnection of public organizations so as fully eliminating bureaucracy, and central control of tax evasion.

In what follows we will describe in a more detailed way the previously mentioned sectors that can be benefited by an ICT investment which can increase efficiency and productivity. The following indicative sectors of the greek economy can potentially be exploited under the best economic terms.

- Exports** In the previous section, we presented the positive trend for Greek exports even during the years of crisis. This means that many Greek companies managed to increase their sales and profits by investing in their future through exports to other countries. This process can be further enhanced and accelerated under a common electronic platform at a national level which will support all firms whose primary turnover comes from abroad. In this platform, all export companies and new companies that have the ability to directly route their production to Exports, through ICT investments would unify processes, bypass bureaucracy, directly issue licenses and directly request funding from Greek banks. Through a unique funding route can achieve minimized costs and time to further support their production lines. All in the same platform that will accelerate the whole process save time, resources and money. Additionally, primary suppliers can be incorporated in this platform (acquiring a unique id) and directly support export companies and production lines with raw materials in a live-time process.
- Agriculture** Agriculture can potentially become a primary pillar for the future development of the Greek economy, specializing in Mediterranean food, in super-foods, and special flavors. Experience shows that ICT investments can accelerate production, speeding-up the introduction of new varieties, and increase the quality of the final products. This can be done through electronic means in order to control diseases, increase productivity of the soil by controlling through electronic sensors humidity, temperature, pH, and other crucial chemical factors, or simply by supervising the fields from air with the assistance of drones. Agricultural firms can collaborate with each other by founding common funds and corporate entities. Later on, these entities can request funding from banks and investors to realize the insertion of technology in production of specific products that have a comparative advantage at a European level. Through economies of scale, can fund their projects with reduced interest rate, increase European market share and increase their total exports activity.
- Tourism** Greece is among the top international destinations with around 33 million arrivals and 17 billion euros income in 2018. This sector has clearly a comparative advantage at a global level. Apart from massive quantities and well known global destinations (i.e., Mykonos, Rhodes, Chalkidiki, Crete, Corfu, etc.), Greece with the assistance of ICT investments will have the ability to focus on boutique, specialized tourism products. This can potentially include geographical areas of unique interest such as lakes, rivers, mountains and cities with archaeological sites or (and most importantly) thematic areas such as conference tourism, religious tourism, food/wine tourism, museum tourism, tourism for people interested in byzantine culture, “philosophy tourism” (i.e., tourism targeting people with deep appreciation for ancient Greek philosophy and culture), and other areas of high interest that Greece can potentially support their development. With the assistance of ICT technologies, a new platform can be created interconnecting boutique hotels and other tourism infrastructure with such niche areas. Through this platform, a new international client will have the ability to view the thematic areas (with the support of videos and other informational material), choose the product and directly make the reservation with the interconnected boutique hotels/resorts. Over the last years, Greece has indeed an upward trend regarding the incoming tourists. What further needed is to focus on this specializing tourism that will add value in economy, create new jobs, and accelerate further economic development.
- Renewable energy** Renewable energy is another field that Greece can further enforce one of its comparative advantages. By exploiting the so called “Greek sun,” the winds of the Aegean, and the current hydroelectric infrastructure, which potentially can be further expanded, Greece has the ability to massively increase its produced energy from renewable resources and covering not only

its needs but also making exports to neighbor countries. Regarding the EU 2020 target (energy consumption from renewable resources), Greece is on track. With the assistance of ICT investments and through electronic platforms, the country can make clever allocations of the produced energy coming from renewable resources. In this way, the country can save money and resources which at a later stage can be re-routed back to the system or moving towards industry and export companies reducing their cost and further increase their productivity and income coming from exports. Most people of Greece are primarily living in two cities: Athens and Thessaloniki. In fact, 70% of Greek people live in the metropolitan areas of these two cities. This fact, in terms of energy consumption, can facilitate the clever energy allocation which with the assistance of ICT investments can potentially lead to economies of scale and the acquiring of benefits mentioned before.

- **Education** Education can potentially become another factor of success for the future economic development of the Greek economy. In this field, Greece can exploit its position in the Balkan Peninsula and its European experience and link this with the European prospects of its neighbors. Right now, many public universities and private colleges have the ability for extroversion and international collaborations. Having already set up international partnerships (especially the private colleges) with British universities and other international academic institutions, Greek private education providers can further expand their activities to Balkan countries by providing distant educational services with the use of ICTs. Electronic platforms for asynchronous or synchronous education can be a fundamental part of this international education dimension. Such products can be directed to individuals, firms that wish to educate their personnel and other private and public entities. This service can add a tremendous and multiplying value in the Greek economy by attracting international distant students and researchers for university and professional education in various fields.

FUTURE RESEARCH DIRECTIONS

We have outlined how ICT investments can foster the Greek economy and for that matter the economy of any country that has similar characteristics with Greece. Once you have a vibrant economy, in which companies strongly use ICT technologies in their business, then one can expect to see new software and hardware companies to flourish. This will further allow companies to create new tools and utilities. In addition, Greek Information Science departments can be associated with such companies preparing their students for their professional life. This of course will have a tremendous multiplying effect in the country and it will stop the brain-drain that happens right now. One can further investigate in full detail all these prospects and abilities.

CONCLUSION

In this chapter we discussed how ICT investments can foster the troubled economy of a country. In particular, we used Greece as a test case to see how its economy can get the required stimulus to begin functioning and maintain a viable rate of economic development. We have examined various economic areas and how these areas can be further expanded with the use of ICT technologies through investments in crucial sectors that appear international comparative advantage. Naturally, there are many other sectors that can be included in such research and analysis.

REFERENCES

- Blanchard, O., & Johnson, D. R. (2012). *Macroeconomics* (6th ed.). Pearson.
- Bodie, Z., Kane, A., & Marcus, A. J. (2013). *Investments* (10th ed.). McGraw-Hill Education.
- Côté, J. E., & Allahar, A. L. (2007). *Ivory Tower Blues: A University System in Crisis*. University of Toronto Press. doi:10.3138/9781442685505
- Kvochko, E. (2013, April 11). *Five ways technology can help the economy*. Retrieved July 12, 2016, from World Economic Forum: <https://www.weforum.org/agenda/2013/04/five-ways-technology-can-help-the-economy/>
- Loverdos, C. K., & Syropoulos, A. (2010). *Steps in Scala: An Introduction to Object-Functional Programming*. Cambridge University Press.
- Papadopoulos, I., & Syropoulos, A. (2018). ICT Investments and Recovery of Troubled Economies. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (pp. 2337–2344). IGI Global. doi:10.4018/978-1-5225-2255-3.ch203
- Staats, F. G. (2012, December). *The Microwork Solution*. Retrieved July 12, 2016, from Harvard Business Review: <https://hbr.org/2012/12/the-microwork-solution>
- Syropoulos, A., Tsolomitis, A., & Sofroniou, N. (2003). *Digital Typography Using LaTeX*. Springer.
- Tsakanikas, A., Danchev, S., Giotopoulos, I., Korra, E., & Pavlou, G. (2014, December). *ICT Adoption and Digital Growth in Greece*. Retrieved July 13, 2016, from http://iobe.gr/press_dtl.asp?EID=59
- United Nations. (2018). Review of Maritime Transport. New York: *United Nations Conference on Trade and Development*.

KEY TERMS AND DEFINITIONS

Economic Crisis: A situation in which the economy of a country experiences a sudden downturn.

Economic Recovery: Is a period of increasing business activity signaling the end of a recession.

Economy: The large set of inter-related production and consumption activities that aid in determining how scarce resources are allocated.

ICT: Is an umbrella term that includes any communication device or application, computer and network hardware and software, satellite systems and so on.

Investment: Is an asset that is purchased with the hope that it will generate income or receive appreciation in the future.

Dimensions of the Digital Divide

Marcus Leaning

University of Winchester, UK

Udo Richard Averweg

eThekweni Municipality, South Africa

INTRODUCTION

The digital divide is a term used to describe a difference in the use of digital (including social) media between and within populations. It concerns the extent to which engagement with digital media causes and is caused by varying demographic factors. The term was first used in 1999 (van Dijk, 2017) though an understanding that a gulf between those able to use and access information through digital media and those who could not, existed since the mid-1990s. The issue received considerable attention in the early to mid-2000s (for example, by the Organization for Economic Cooperation and Development (OECD)) yet in contemporary times the term seems to have drifted out of the public consciousness. However, there are still large differentials between regions of the world in terms of rates of access to digital media. There are also notable differences between communities in developed countries in terms of what forms of access people have and what they are able to do with this access.

As understanding of the divide has progressed conceptualizing the various reasons people are not able to avail themselves of the potency of computers and the internet has become more sophisticated. In this article, the authors contend that the idea of the digital divide can be understood to operate in a range of different ‘dimensions’. Previous work on the topic has identified different historically situated ‘orders’ of the digital divide and in some ways the dimensions under consideration here match these. However, such accounts tend to locate the forms of divide as occurring sequentially and that the latest form of divide presents the most prescient problems. In contrast the authors contend that the differing forms are still present and that the problem is a multi-faceted one which requires a multi-level approach to address. In this chapter the authors consider the three main dimensions that impact upon a person’s ability to make use of digital media.

BACKGROUND

In its simplest terms the digital divide refers to a form of social stratification that is simultaneously enacted and furthered by an individual’s ability to utilize digital media to render their own self-interest (Leaning, 2017a). That is, our access to and use of digital media in part determines our social opportunities but is simultaneously related and determined by forms of social inequality (Ragnedda, & Mu, 2013). It impacts an individual’s ability to use digital media to assist them in their lives but also impacts upon their engagement with political processes. Accordingly, the digital divide is both a problem for individual people and also for the operation of democracy and citizenship.

In this article the authors first consider the physical (access to the equipment to connect) and material (ability to afford the expense of connection) digital divide and rates of access across and within a

DOI: 10.4018/978-1-7998-3479-3.ch115

number of countries. Second, the authors then look to the issues of training and education that impact upon people's ability to access computing technology. Third, the authors look at what people are actually doing online and will note that demographic differences between people are often also manifest in their forms of behavior and ability to leverage digital communications to their advantage.

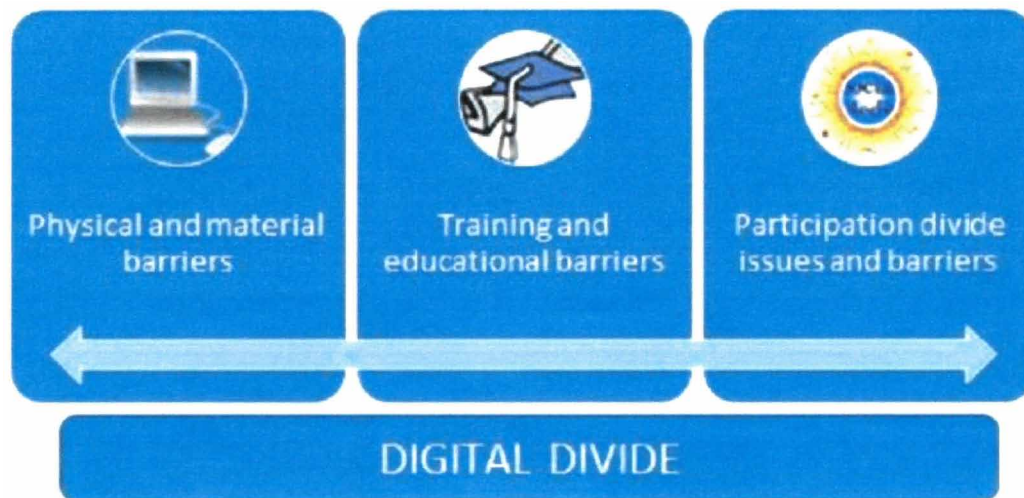
The authors term these different barriers as different dimensions to highlight their interrelatedness and concurrency. It is felt that to refer to the problems as barriers or orders implies that the issues occur sequentially. While they were identified sequentially, treating them in this manner means that the complexity of the issues and the interrelatedness of the three issues may not be fully appreciated. Instead the authors refer to the issues as dimensions of a complex problem. The dimensions the authors identify here operate in concert and thus cannot be addressed in a piece-meal or singular fashion.

FOCUS OF THE ARTICLE

This article will look at different aspects of the digital divide. The article is structured so that the broad themes (different dimensions) cover the following aspects, as reflected in Figure 1:

- physical and material barriers;
- training and educational barriers; and
- participation divide issues and barriers.

Figure 1. Dimensions of the digital divide in contemporary society



The authors now explore and discuss these each of these different dimensions in the hope that readers may have their awareness and understanding of these digital divide topics strengthened.

Physical and Material Barriers

There are now high levels of access to the internet in most developed countries and indeed across other regions of the world. For example, in North America and the Caribbean the average access rate is 76.2%, in Europe it is 80.2%, in South America it is 65.3%, across Asia and the Middle East access is 47.0%, in Europe it is 80.2%, in Australia and Oceania it is 69.6%, while in Africa it is as low as 47.0% (Leaning, 2017b). There are, however, very significant disparities of access and many countries lag far behind the most connected regions: while Australia has nearly 85.0% of its population online, only 7.0% of Togo's population has the same access. It must be noted, however, that for many access to the internet comes not through a connected computer but through a mobile device with an internet subscription. Such practices further complicate the picture as in many parts of the world users may possess multiple subscriptions to take advantage of cheaper rates for certain actions.

Furthermore, within countries there are further divides, these replicate and indeed exacerbate other forms of social inequality. Forms of division such as gender (Cooper, 2006; DiMaggio, Hargittai, Neuman, & Robinson, 2001; Dixon, Correa, Straubhaar, Covarrubias, Graber, Spence, & Rojas, 2014), educational level (van Dijk, & Hacker, 2003), ethnicity and race (Hoffman, Novak, & Schlosser, 2001; Jackson, Zhao, Kolenic III, Fitzgerald, Harold, & Von Eye, 2008), language (Gurstein, 2003; Mallikarjun, 2004), social class and financial standing (Clayton, & Macdonald, 2013), age (Cresci, Yarandi, & Morrell, 2010), sub-national (Chen & Wellman, 2004) and intra-national regions (Vicente, & López, 2011) have been considered as being important in determining an individual's ability to access computers and internet technology. Such lack of access for members of these groups often results in inequality being further entrenched. Access to the internet is both restricted by various forms of social inequality but also exacerbates and contributes to such inequality as the possible benefits of digital media are denied.

Attempts to deal with these problems run into multiple difficulties. van Dijk's 2005 study determines that a user's motivation may prove to be a barrier that exists prior to access. Of those not online, a small percentage are in that position as they lack the motivation to be online – they are 'want nots' as opposed to 'have nots' and the reasons for not wishing to be online are complex. This may be because of a fear or anxiety, a lack of time, not seeing the value of access, fears over the effects of computing or simply a disinterest in computers. One needs to be cautious about saying such beliefs are problems, however, and there may be many legitimate reasons for not wanting to be online. Once the barrier of motivation is surpassed, the key barrier facing users has been understood to be one of access.

A number of academics (van Dijk, & van Deursen, 2014; Warschauer, 2004) have noted that the issue of access to computers and the internet forms is more complicated than the issue of having direct physical access. van Dijk (2005) and Dijk & van Deursen (2014) note access is comprised of two separate aspects: physical access – which consists of the direct contact with an internet enabled computer – and material access – the wealth to be able to afford the expenses of being online such as broadband subscriptions, costs of various services and 'apps' and subscriptions to services. If one considers the use of mobile devices in developing countries, a number of services also require the user to have credit cards and other financial tools to avail themselves of access - barriers that prevent those of limited wealth from engaging. While many may have physical access through using a computer, through work or in cybercafes or other locations or even possessing one material access proves a more difficult barrier to overcome as it is linked to financial inequality. Warschauer (2004) notes the complexity of successful computer usage "access to ICT is embedded in a complex array of factors encompassing physical, digital, human, and social resources and relationships".

During the late 1990s and much of the early 2000s the issue of the digital divide received attention at the highest of governmental, non-governmental, corporate and charitable levels (Klein, 2004) as well as extensive academic interest (Norris, 2001). Solving the problem of the digital divide became a significant area of domestic and international development activity and consumed large amounts of funding. Many of the solutions proffered involved the greater deployment of technology allowing individuals who were not connected the opportunity to become connected. While in developed countries the issues of physical and material access have been ameliorated to a large degree, they still present significant issues to people in the developing world and are a very current problem.

Training and Educational Barriers

Significant effort was put into addressing physical and material barriers during the early 2000s. Though there were a few instances, such as Costa Rica, where direct state investment was used successfully to provision digital service, in most cases the approach to provision has primarily been through the extension of commercial services be they commercial broadband or mobile services. However, it soon became evident that access alone was not the entire solution and the simplistic binary nature of the divide began to be questioned. This resulted in more sophisticated models being developed to account for the digital divide.

The new models conceptualized the digital divide not in binary terms but instead saw a spectrum of ability to access and use of networked computers (van Dijk, & Hacker, 2003; Warschauer, 2002, 2004). In such accounts, though access was still a barrier, other factors became increasingly recognized as preventing full engagement with digital media (Hargittai, 2002) and Warschauer's (2004) critique that access was part of a complex range of relationships and social factors was developed further by a number of authors. The most significant problem identified in these accounts was the lack of skills in using computers and internet technologies and this soon became the focus of attention in efforts to address the divide.

A number of new models of skilled usage were proposed to account for differences in participation (Barclay, & Duggan, 2008; van Dijk, 2005). These models determined a deficit in skills as being central to restricting people availing themselves of the benefits of digital media in terms of activities such as economic activity (König, Lorenz Graf-Vlachy, & Mammen, 2016), business (Arendt, 2008), health (Norman, & Skinner, 2006) and political engagement (Morris, & Morris, 2013).

Actual Skills Which Need To Be Improved

van Dijk (2005; van Dijk & van Deursen, 2014) advanced the approach further and prepared exhaustive methods for determining the actual skills that needed to be improved so that the divide could be addressed. van Dijk makes use of a model of capital - initially proposed in the work of French sociologist Pierre Bourdieu (Bourdieu argued that social inequality was not simply a matter of not having enough money. Rather he identified three forms of capital (financial, social and cultural) that facilitated social stratification. To these three, the Dutch critic Cees J Hamelink (2000) proposed that as the Information Age progresses, a fourth form of capital must be added, that of information skills as this assists a person's ability to progress in society).

van Dijk & van Deursen (2014) develop this approach and offer a six-part model of the skills required by individuals to bridge the divide. This model has proven significantly influential in developing numerous programs to equip students with the skills for using digital media in contemporary times and

it is worth considering the model in some detail. Van Dijk & van Deursen contend that users need two new sets of skills specific to the digital age:

- *operational skills* – the practical know how to engage with digital media; and
- *formal skills* – the technical practices for using hyper media.

Van Dijk then identifies four further content related skills:

- *information skills* – the basic skills of handling information;
- *communication skills* - being able to communicate on the internet;
- *creative skills* - for digital media content production; and
- *strategic skills* – to use one's creative and information skills in a useful manner.

However, while van Dijk & van Deursen's model presents a useful tool for evaluating and developing educational practices for addressing the second dimension of the digital divide he does not advance the nature of the critical approach to information that is developed. Though there is some very brief mention of evaluating information, the proposed model does not

- incorporate any real consideration of how texts themselves should be considered;
- consider the ways in which information that is mediated through a variety of sources and often relayed through different social media lens and channels; and
- explore how a user's data is used by social media platforms and other organisations.

Thus the skills identified by van Dijk & van Deursen and incorporated into many models, also need development and updating to address more recent activities such as the use of individuals social media data profiles. One potential option to address this is the development of a new set of competencies broadly referred to as Media and Information literacy. This literacy incorporates a more critical aspect and utilizes some of the critical methods that have been present in media education for some years but not previously applied to information resources.

A further concern with the information-intensive model developed by van Dijk is that while some attention is paid to the creation of digital media content in the model, this is rather limited in its understanding of the nature of participatory culture and the differing levels of engagement with digital media by different communities.

Participation Divide Issues and Barriers

The development of extensive forms of training provision and skilling practices is understood to have addressed the problems identified in the second dimension of the digital divide. Computing and information skills became an accepted part of the educational curricula for students of all ages in many countries. Moreover, the growing acceptance that people regularly need to update their skills in the workplace and the emergence of the idea of lifelong learning resulted in numerous courses for adults outside of traditional educational spaces. Libraries and similar venues became recognized as places to equip people with the necessary skills to use digital media.

Alongside these changes in how education was understood, during the mid-2000s there was also a range of technological developments (such as the development and proliferation of mobile devices and

the emergence of web 2.0 and social media platforms). There were also other developments and systems which afforded users a greater opportunity to produce and disseminate their own content.

The ability to perform such actions was investigated by academic researchers with one branch of research considering the demographic characteristics of those who were producing and sharing material. This area was expressly concerned with the issue of whether (if access and training were equal) an individual's propensity to engage in the productive aspects of social media is adversely affected by 'real world' material, cultural and non-technical educational advantages – are extant inequalities reduced or exacerbated by an individual's ability to participate? This approach was specifically concerned with whether social inequality can be addressed and reduced through digital media, whether social media simply reflects extant social inequalities or whether media further worsens existing problems.

However, establishing a direct link between markers of social inequality and a propensity to engage with social media and participatory culture is difficult. Micheli (2015) finds no correlation between the parental employment status of users and their proclivity to engage with social media. However, a link is noted between particular activities – such as information seeking - which contributes to the building of social capital and the parental social class. Moreover, engaging in social media is a varied activity ranging from a largely consumptive activity through to a highly productive one in which users create content – social media can be and is put to many different uses in people's lives (Lutz, & Hoffmann, 2014). Furthermore, Blank (2013) and Blank & Reisdorf (2012) note that content production is not a single activity but incorporates many different practices, ranging from posts on social networks, creating written texts, static and video texts, acts of citizen journalism and editing, and repurposing other's content.

The relationship of social inequality and the propensity to participate is complex. Once issues of access and training are addressed, a user's direct ability to use social media and the amount of use of social media usage does not seem affected by the status or forms of social disadvantage their users endure. Indeed, in some instances those from disadvantaged backgrounds have a greater degree of participation than those from a non-disadvantaged social milieu. Blank (2013) and Hoffmann, Lutz & Meckel (2015) both noted that a larger proportion of time is spent on social media by those from low socio-economic backgrounds than those from more wealthier groups. However, what forms of social media are engaged in do seem differentiated by levels of educational attainment and parental social class (in the case of younger users).

Brake (2014), in drawing upon a data set from Oxford University and using a broad range of literature, notes how activities of a more creative nature such as blogging, creating content on Wikipedia or forms of citizen journalism are activities engaged in more by those of higher economic status than those of a lower one. Similarly Hargittai & Jennrich (2016) and Hargittai & Walejko (2016) note how socio-economic status is linked to the production of creative content. Correa (2016) notes no difference in the amount of social media use between members with different educational levels when the use is for social purposes but does identify a difference between users when the task is searching for information. Thus it appears there are particular productive activities that are performed more by people of a higher socio-economic group than those from lower socio-economic groups.

While certain proponents of digital and participatory culture contend that use of any social media builds communication skills and confidence (Jenkins, 2006), it appears there are distinct activities that seem to be performed by particular groups. Some of these activities, such as searching for information, using social media for instrumentally career-focused networking and building personal brands, are correlated with users of higher economic standing and other forms of privilege. However, it is difficult to determine the direction of travel – do socially advantaged groups use social media because they are advantaged or are they advantaged because they use social media in this way? A number of studies (see,

for example, Radovanović Hogan, & Lalić, 2015) indicate that the social status impacts upon the way in which social media is used. The determination of which social media practices a user engages with appeal is, to a degree, determined by the level of social advantage a person has. Social media usage is a consequence of prior existing social factors rather than a mitigating factor to them.

SOLUTIONS AND RECOMMENDATIONS

The various dimensions of the digital divide pose a challenging problem. Initial efforts were made to alleviate the disadvantages faced by certain groups and populations which would exclude them from obtaining the advantages of access to digital media. The recognition that education and training were a further requirement to address the issue added a further aspect to the problem and also refocused efforts away from the physical and material problems onto human and social factors. The targeted educational practices of the various digital literacy programmes developed to address the skills divides are still very current issues.

A raft of new initiatives are being developed by international agencies (such as UNESCO) to broaden the training beyond the primarily technical and instructive methods into more nuanced, critical and sophisticated approaches. These new approaches seek to expand training beyond the ideas of how to use communicative technology into challenging the problems of 'Fake News' and propaganda.

The growth of social media also posed questions about how socially disadvantaged groups used different forms of social media and engaged in different practices. This aspect of the digital divide is significantly problematic as research seems to indicate that the expansion of the popularity of social media has resulted in it actually becoming a new space in which extant disadvantage can be enacted. Accordingly, the provision of access to and skilled use of digital media technology and social media may not be a means to alleviate social disadvantage but a new channel by which such disadvantage is continued.

The provision of services, delivery of training and even developing the adoption of specific practices may have side-effects to the intended alleviation of disadvantage and instead social media becomes a further arena through which social inequality is enacted. The provision of such technology should not be seen as a means to alleviate inequality but may indeed be a new means by which such inequality is enacted, facilitated and accelerated (Leaning, 2009).

Accordingly, how one understands the digital divide may be understood not as one of multiple barriers to overcome but rather as one of a social problem of which there are multiple dimensions.

FUTURE RESEARCH DIRECTIONS

To advance understanding of the digital divide, the authors advocate that future research be directed towards a greater understanding of how existing social disadvantages are enacted through social media - explicitly through the differences in use by differing populations and how certain practices can assist in ameliorating social disadvantage. Such digital divide research should dovetail with the approaches currently undertaken in OECD countries – see www.oecd-ilibrary.org

CONCLUSION

The three main dimensions that the digital divide can be understood to operate in were articulated so that the associated prescient problems may soon gain resolution. The authors have shown that in contemporary society the digital divide is a multi-faceted problem which requires a multi-level approach to address the dimensional issues raised.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Some of this text appeared in a three-part article series in the *Journal of the Southern Africa Institute of Management Services (SAIMAS)*. Permission was granted by the Editor of SAIMAS to reuse this text.

REFERENCES

- Arendt, L. (2008). Barriers to ICT adoption in SMEs: How to bridge the digital divide? *Journal of Systems and Information Technology*, 10(2), 93–108. doi:10.1108/13287260810897738
- Barclay, C., & Duggan, E. (2008). Rethinking the Digital Divide: Towards a Path of Digital Effectiveness. *41st Annual Hawaii International Conference on System Sciences*. 10.1109/HICSS.2008.376
- Blank, G. (2013). Who creates content? Stratification and content creation on the Internet. *Information Communication and Society*, 16(4), 590–612. doi:10.1080/1369118X.2013.777758
- Blank, G., & Reisdorf, B. C. (2012). The Participatory Web. *Information Communication and Society*, 15(4), 537–554. doi:10.1080/1369118X.2012.665935
- Brake, D. (2014). Are We All Online Content Creators Now? Web 2.0 and Digital Divides. *Journal of Computer-Mediated Communication*, 19(3), 591–609. doi:10.1111/jcc4.12042
- Chen, W., & Wellman, B. (2004). The global digital divide—within and between countries. *ITandSociety*, 1(7), 39–45.
- Clayton, J., & Macdonald, S. J. (2013). The limits of technology: Social class, occupation and digital inclusion in the city of Sunderland, England. *Information Communication and Society*, 16(6), 945–966. doi:10.1080/1369118X.2012.748817
- Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning*, 22(5), 320–334. doi:10.1111/j.1365-2729.2006.00185.x
- Correa, T. (2016). Digital skills and social media use: How Internet skills are related to different types of Facebook use among ‘digital natives’. *Information Communication and Society*, 19(8), 1095–1107. doi:10.1080/1369118X.2015.1084023
- Cresci, M. K., Yarandi, H. N., & Morrell, R. W. (2010). The digital divide and urban older adults. *Computers, Informatics, Nursing*, 28(2), 88–94. doi:10.1097/NCN.0b013e3181cd8184 PMID:20182159

- DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the Internet. *Annual Review of Sociology*, 27(1), 307–336. doi:10.1146/annurev.soc.27.1.307
- Dixon, L. J., Correa, T., Straubhaar, J., Covarrubias, L., Graber, D., Spence, J., & Rojas, V. (2014). Gendered space: The Digital divide between male and female users in internet public access sites. *Journal of Computer-Mediated Communication*, 19(4), 991–1009. doi:10.1111/jcc4.12088
- Gurstein, M. (2003). Effective use: A community informatics strategy beyond the digital divide. *First Monday*, 8(12). Advance online publication. doi:10.5210/fm.v8i12.1107
- Hamelink, C. J. (2000). *The Ethics of Cyberspace*. Sage Publications.
- Hargittai, E. (2002). Second-level digital divide: Mapping differences in people's online skills. *First Monday*, 7(4). Advance online publication. doi:10.5210/fm.v7i4.942
- Hargittai, E., & Jennrich, K. (2016). The Online Participation Divide. In *The Communication Crisis in America, And How to Fix It* (pp. 199–213). New York: Palgrave Macmillan, US. doi:10.1057/978-1-349-94925-0_13
- Hargittai, E., & Walejko, G. (2008). The Participation Divide: Content creation and sharing in the digital age1. *Information Communication and Society*, 11(2), 239–256. doi:10.1080/13691180801946150
- Hoffman, D. L., Novak, T. P., & Schlosser, A. E. (2001). The evolution of the digital divide: Examining the relationship of race to Internet access and usage over time. *The Digital Divide: Facing a Crisis or Creating a Myth*, 47–97.
- Hoffmann, C. P., Lutz, C., & Meckel, M. (2015). Content creation on the Internet: A social cognitive perspective on the participation divide. *Information Communication and Society*, 18(6), 696–716. doi:10.1080/1369118X.2014.991343
- Jackson, L. A., Zhao, Y., Kolenic, A. III, Fitzgerald, H. E., Harold, R., & Von Eye, A. (2008). Race, gender, and information technology use: The new digital divide. *Cyberpsychology & Behavior*, 11(4), 437–442. doi:10.1089/cpb.2007.0157 PMID:18721092
- Jenkins, H. (2006). *Fans, Bloggers, and Gamers: Exploring Participatory Culture*. New York University Press.
- Klein, H. (2004). Understanding WSIS: An Institutional Analysis of the UN World Summit on the Information Society. *Information Technologies and International Development*, 1(3-4), 3–13. doi:10.1162/1544752043557341
- König, A., Lorenz Graf-Vlachy, D., & Mammen, J. (2016). Second-Order Digital Inequality: The Case of E-Commerce. *Mechanisms driving technology use in the context of digital inequality: A series of essays on the role of social influence, socio-cognitive processes, and socio-economic determinants*, 228.
- Leaning, M. (2009). *The Internet, Power and Society: rethinking the power of the internet to change lives* (1st ed.). Chandos. doi:10.1533/9781780631684
- Leaning, M. (2017a). *Media and Information Literacy: An Integrated Approach for the 21st Century*. Elsevier. doi:10.1016/B978-0-08-100170-7.00001-9

Leaning, M. (2017b). Internet Accessibility: Continental Comparison. *UNESCO 7th Media and Information Literacy and Intercultural Dialogue Conference*.

Lutz, C., & Hoffmann, C. P. (2014). *Towards a broader understanding of the participation divide(s)*. Paper presented at the ICA Preconference Communication and “The Good Life” - Around the World After Two Decades of the Digital Divide, Seattle, WA.

Mallikarjun, B. (2004). Indian multilingualism, language policy and the digital divide. *Language in India*, 4(4).

Micheli, M. (2015). What is New in the Digital Divide? Understanding Internet Use by Teenagers from Different Social Backgrounds. In L. Robinson, S. R. Cotten, J. Schulz, T. M. Hale, & A. Williams (Eds), *Communication and Information Technologies Annual* (pp. 55-87). Bingley, UK: Emerald Group Publishing Limited. doi:10.1108/S2050-206020150000010003

Morris, D. S., & Morris, J. S. (2013). Digital Inequality and Participation in the Political Process Real or Imagined? *Social Science Computer Review*, 31(5), 589–600. doi:10.1177/0894439313489259

Norman, C. D., & Skinner, H. A. (2006). eHealth literacy: Essential skills for consumer health in a networked world. *Journal of Medical Internet Research*, 8(2), e9. doi:10.2196/jmir.8.2.e9 PMID:16867972

Norris, P. (2001). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. Cambridge University Press. doi:10.1017/CBO9781139164887

Radovanović, D., Hogan, B., & Lalić, D. (2015). Overcoming digital divides in higher education: Digital literacy beyond Facebook. *New Media & Society*, 17(10), 1733–1749. doi:10.1177/1461444815588323

Ragnedda, M., & Mu, G. W. (2013). *Introduction. The Digital Divide: The internet and social inequality in international perspective*. Routledge. doi:10.4324/9780203069769

van Dijk, J., & Hacker, K. (2003). The Digital Divide as a Complex and Dynamic Phenomenon. *The Information Society*, 19(4), 315–326. doi:10.1080/01972240309487

van Dijk, J. A. (2005). *The deepening divide: Inequality in the information society*. Sage Publications.

van Dijk, J. A., & van Deursen, A. J. A. M. (2014). *Digital Skills*. Unlocking the Information Society. doi:10.1057/9781137437037

van Dijk, J. (2017). Digital Divide: Impact of Access. In P. Rossler (Ed.), *The International Encyclopedia of Media Effects*. Wiley.

Vicente, M. R., & López, A. J. (2011). Assessing the regional digital divide across the European Union-27. *Telecommunications Policy*, 35(3), 220–237. doi:10.1016/j.telpol.2010.12.013

Warschauer, M. (2002). Reconceptualizing the Digital Divide. *First Monday*, 7(7). Advance online publication. doi:10.5210/fm.v7i7.967

Warschauer, M. (2004). *Technology and Social Inclusion: Rethinking the Digital Divide*. MIT Press. doi:10.7551/mitpress/6699.001.0001

ADDITIONAL READING

Cooper, J., & Weaver, K. D. (2003). *Gender and computers: Understanding the digital divide*. Psychology Press. doi:10.4324/9781410608932

IBM-Research. (2011). IBM Next 5 in 5 2011: Mobile. <https://www.youtube.com/watch?v=tuisda1q6ns>

Jenkins, H. (2009). *Confronting the challenges of participatory culture: media education for the 21st century*. MIT Press. doi:10.7551/mitpress/8435.001.0001

Min, S.-J. (2010). From the digital divide to the democratic divide: Internet skills, political interest, and the second-level digital divide in political internet use. *Journal of Information Technology & Politics*, 7(1), 22–35. doi:10.1080/19331680903109402

Pick, J., & Nishida, T. (2015). Digital divides in the world and its regions: A spatial and multivariate analysis of technological utilization. *Technological Forecasting and Social Change*, 91, 1–17. doi:10.1016/j.techfore.2013.12.026

Rhoades, H., Wenzel, S., Rice, E., Winetrobe, H., & Henwood, B. (2017). No digital divide? Technology use among homeless adults. *Journal of Social Distress and the Homeless*, 26(1), 73–77. doi:10.1080/10530789.2017.1305140 PMID:31097900

Selwyn, N. (2004). Reconsidering Political and Popular Understandings of the Digital Divide. *New Media & Society*, 6(3), 341–362. doi:10.1177/1461444804042519

Straubhaar, J., Spence, J., Tufekci, Z., & Lentz, R. G. (2012). *Inequity in the technopolis: Race, class, gender, and the digital divide in Austin*. University of Texas Press.

KEY TERMS AND DEFINITIONS

1st Dimension Divides: The physical and material barriers to participation.

2nd Dimension Barriers: The educational and training barriers to participation.

3rd Dimension Barriers: The differences in levels of participation.

Critical Information Literacy: The higher-level skills and understanding that result in a person being able to make use of digital media to further their own interests.

Digital Participation: The idea that full engagement with digital opportunities assists in furthering one's self interest.

Information and Information Literacy: The skills needed to successfully use computing technology.

Social Stratification: The way in which a society is organized so that opportunities and benefits are more available to certain members of a society than others.

Importance of Digital Literacy and Hindrance Brought About by Digital Divide

13

Mohammad Izzuddin Mohammed Jamil

Universiti Brunei Darussalam, Brunei

Mohammad Nabil Almunawar

 <https://orcid.org/0000-0001-5296-2576>

Universiti Brunei Darussalam, Brunei

INTRODUCTION

In this century we have seen rapid advancements in technology, and more innovative devices spawn every year. For better or for worse, technology have changed the way people go about their everyday lives. Humanity have achieved technological feats that were previously thought impossible such as computer processing and online communication. Along the way, a technological paradigm has emerged that has further altered the way people live; digital technology. Digital technology includes all types of devices and electronic applications that use and store information in the form of binary code, and these information can usually be altered by capable devices such as desktops, laptops, tablets and smartphones. Presently, information is available in vast amounts on a global computer network called the Internet. One cannot simply utilise technology, it requires a certain degree of skills and competencies. These skills and competences are commonly known as *digital literacy*. Digital literacy showcases an individual's grammar, vocabulary, writing, and typing skills on various platforms, and the appearance of the Internet means that digital literacy also extends to the capability of individuals in extracting and critically analysing information on the Internet and social media. With these advancements in the field of technology, it is vital to understand the importance of digital literacy.

People who possess sufficient amount of digital literacy will be able to reap the benefits of digital technology and will have their lives continuously improved. However, not everyone experiences the same benefits. A global phenomenon called digital divide has recently appeared, whereby there exists a gap between those who have access to technology and those without. Across the globe, Internet technology has developed unevenly throughout. Countries who have access to the Internet are able to develop, while countries who have no means of access or limited access to technology and the Internet are left behind socially and economically, and remains poor.

Thus, it is the objective of this article to provide a clear and concise definition of digital literacy. While many studies have explained in detail the definition of digital literacy, this chapter would like to complement that definition by applying it in the context of an Input-Process-Output (IPO) model, in order to provide more clarification and to help the general public better understand the concept of digital literacy. The article would then prioritise its objective to justify why having a certain degree of literacy is important from different industries. The importance of digital literacy is even more pronounced with the rapid rise in the usage of Internet (Internet World Stats, 2018, Dec 31). However, digital divide has plagued the world since the Internet first came into existence. Due to the negative social and economic

DOI: 10.4018/978-1-7998-3479-3.ch116

implications of digital divide, the article would then suggest solutions and recommendations on ways to tackle digital divide.

BACKGROUND

The word *digital literacy* and the topic surrounding it is relatively new, going back only as far as the late 1990s (Lanham, 1997; Pool, 1997). Interestingly, digital literacy existed years after the Internet first made its appearance in 1991 (Berners-Lee, Cailliau, Groff & Pollermann, 1992). Since then, one of the earliest literature introducing digital literacy is done by Gilster (1997), in which he provided the explanation for the definition of digital literacy as possessing the ability to process information via technological means. However, the broad nature of this definition may have caused more questions than answers. The preceding literature that follows have attempted to interpret, analyse and evaluate this definition (Bawden, 2001; Robinson & Bawden, 2001, June; Williams, 2006). Bawden (2008) explained Gilster's definition further by applying it in the context of the present time; the digital age. The Input-Process-Output (IPO) model was first used to evaluate training (Bushnell, 1990), and has since been used in literature to describe a particular programming process (Curry, Flett & Hollingsworth, 2006; Grady, 1995; Goel, 2010; Zelle, 2004), but none have covered in the context of digital literacy.

Digital literacy is not to be confused with other similar literacies (Koltay, 2011) such as computer literacy (Hoffman & Blake, 2003), media literacy (Livingstone, 2004), and information literacy (Bruce, 1999), whose words have existed long before the emergence of digital literacy (Bawden, 2001). The existence of these words is in line with technological advancement; the more technology advances, more literacies begun to appear. Albeit the literacies are used interchangeably, ultimately they possess slightly different meaning from one another.

With technology progressing at a rapid pace, more studies have begun to appear in order to examine the underlying meaning behind digital literacy (Buckingham, 2008; Martin, 2008). Digital literacy requires skills and training to be able to competently process and produce information on digital media, which is far more easier for younger generations to master as compared to senior generations. In today's digital age, technology have made lives more convenient and as a result, people have started to rely on technology and are dependent on it to get things done. These studies culminated in the realisation on the importance of having a certain degree of digital literacy in order to survive today's digital age (Eshet, 2004; Martin, 2005).

Apart from digital literacy, digital divide is also an area that only been recently explored (Hoffman & Novak, 1998; Sidney Howland, 1998). Recent researches on digital divide include the study of the causes of digital divide in different context such as in tourism industry (Minghetti & Buhalis, 2010), as well as the context of a group of countries such as European (Union Cruz-Jesus, Oliveira & Bacao, 2012). In fact, the usage of the word digital divide coincides not long after literatures in digital literacy have begun to appear (Jaeger, Bertot, Thompson, Katz & DeCoster, 2012; Warschauer, 2002). As a result, there are currently only a few studies that determines how to solve digital divide (Ali, 2011; Lopez, Gonzalez-Barrera, & Patten, 2013).

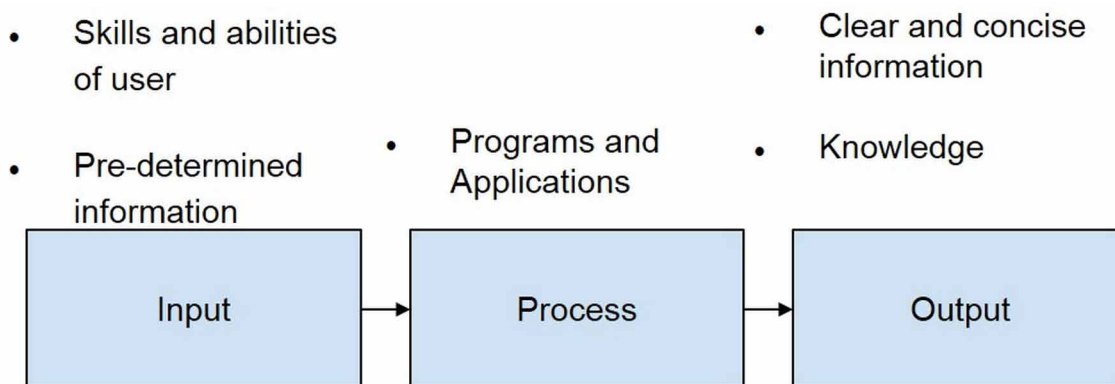
FOCUS OF THE ARTICLE

Definition of Digital Literacy

This paper would define the concept of digital literacy as the skills, abilities and competence of each human individual in using technology or any electronic devices with Information, Communication and Technology (ICT) capability (Buckingham, 2010), in order to discover, assess and create information by way of virtual or digital means (Buckingham, 2015). This means that, an individual must possess a certain degree of knowledge and mastery in ICT to properly use that information. Individuals who are digitally literate would then be able to use the information appropriate to their needs by constructing and interpreting meaning in the form of texts, sounds and images.

Deriving from the widely used model to describe how a typical information processing system operates (Bushnell, 1990; Rabitz, Aliş, Shorter & Shim, 1999), this paper seeks to adapt the Input-Process-Output model (IPO) model in the context of Digital Literacy. Because Digital Literacy involves the use of similar information processing system to generate the required output, it is possible to adapt the model in the context of digital literacy as the IPO model embodies how the inputs are received from a source, whereby the inputs are processed by the processing system, which produces outputs in the form of results. In the study of Computer Science, the process is highly similar to a computation of mathematical expressions, whereby a computer system is used to equate raw mathematical objects into useful mathematical expressions (Al-Rfou, Alain, Almahairi, Angermueller, Bahdanau, Ballas, & Bengio, 2016). The nature of Digital Literacy involves the transition and transformation of information while the IPO model serves as a program that can showcase and embody how that information are undergoing the transformation. The IPO model seeks to specifically identify the items in question for input, process and output.

Figure 1. Input-Process-Output (IPO) Model in the concept of Digital Literacy



To further apprehend the underlying meaning behind digital literacy, it is important to further understand the Input-Process-Output (IPO) model (Ilgen, Hollenbeck, Johnson & Jundt, 2005) with the use of examples. The IPO model forms the basis of many frameworks including digital literacy. From the IPO model of Figure 1, skills and competences of users are considered the input. Input also includes the critical thinking skills of users, in which how capable they are in making evaluation and critical judge-

ments with the information in hand. Pre-determined information from sources such as the Internet and social media also serves as input.

At the next stage, process takes into account software and programs in computers that are being used to create and process new information. Commonly-used applications such as word processor programs, spreadsheet programs, presentation programs, and database management system are some examples. These programs simply enable the altering of information, and made the task of altering information easier. However, how the information is altered depends on the skills of users.

The model then produces output in the form of processed information. This new information created by the application in the process stage can also be synthesised with the aforementioned pre-determined information from the input process. Processed information must be as clear as possible in order to be understood by the majority of people. Output also occurs in the form of contextualized information called knowledge. While information is raw data converted into a meaningful context, knowledge is insights and understanding that would benefit an individual (Kanehisa et al., 2013).

Digital literacy is not to be confused with the general definition of technology itself. Technology include crude and physical tools that makes tasks easier, and these tools may have no digital features whatsoever. The more accurate term for technology associated with digital literacy is digital technology. While digital literacy initially focuses on digital skills and stand-alone computers, it has since shifted to network computers such as the Internet and social media.

Importance of Digital Literacy

For this subheading, the significance of possessing adequate digital literacy would be discussed from the various sectors in a typical economy, as this chapter argues for a push towards digital literacy among the populace across the globe. This includes broader perspectives such as the public sector and the simple necessity to survive in today's digital economy, down to more specific point of views which include Classroom Education, Retailing and Healthcare. In some case, current examples would be made in each instance as well as references to IPO model in the various sectors.

To Survive Today's Digital Society, and Foster Autonomy Rather Than Dependence

Digital literacy must no longer be seen as an option for individuals, but rather a necessary aspect in daily lives. Throughout the world in the 21st century, almost every aspect runs on digital technology. Consider, the smartphones are used for simple communication and social networking activity, while computers are being used as a platform for web-browsing, to advanced video, audio and image programming. Technology has become a permanent fixture in our lives. With example from the IPO model, individuals must have the ability to comprehend and evaluate instructions in the form of texts on graphical user interfaces, and then create meaningful context out of it (Eshet, 2004). However, there is often the issue of ethics in digital context such as copying other people's work on the Internet, as well using the information displayed to create something that is deemed morally wrong (Capurro, 2009). Thus, it is vital that a boundary be established for individuals to differentiate between right and wrong in the form of digital legislations. For instance, due to China having one of the highest number of Internet users in the world, China has implemented digital legislations in order to reduce the rate of cybercrimes (Qi, Wang, & Xu, 2009).

The consequence is that society have begun to be dependent on technology and thus, cannot live without it. This results in the notion that we are living in an age of digital, rightly called *digital society* (Martin, 2008). How would an individual expect to survive in the digital society if they do not have a single clue on how to utilise digital technology, when the completion of their tasks hinges on it. Thus, in order to adapt to this technological paradigm, it is essential that each and every individual be digitally literate to survive. However, there must be an establishment between the boundary of dependence and autonomy. With digital literacy, an individual can properly learn that digital technology are not to be dependent on completely, rather it is something that can support an individual be more autonomous and self-independent in order to better themselves in terms of surviving in today's society (Kipnis, 2012).

Classroom Education

Digital literacy is of the utmost importance in the realm of education. Traditionally, teachers use whiteboards and chalkboards to explain a subject matter, and print exercises in the form of hardcopy for students to complete. These crude methods however are slowly becoming obsolete, as students are accustomed to using their smart devices in class. Fortunately, digitally literate teachers and students can take advantage of the conveniences of these digital technology (Rhodes & Robnolt, 2009). Students can showcase their learning by understanding how to use Internet features such as search engines, web browsers, blogs and emails. Teachers and students are able to create online content to be utilised both in and out of the classroom, without having to wait for the next class. These online contents tend to include texts, images, sounds, videos, and animation, which makes learning more engaging. In class, a new interactive whiteboard can be used on a daily basis. Without digital literacy, teachers and students are not able to take advantage of digital medium. The term for this resourcefulness is E-Learning (Driscoll, 2002).

Outside the classroom, students are often left to their own devices to do their homework, and in some cases, have no means to contact their teachers. Any enquiry has to be made in the next class, and timetables are often so fixed that they are often not flexible enough to change according to students' availability. Whereas in the old days teachers could only get in touch with students in class, teachers can now connect with their students via social media. Smartphones have enabled teachers to track the progress of students and update them of any changes in timetable, while enquiry only requires a few button pushes for students. Outside of education, it is a means for teachers to observe the activities of their students. Thus, digital literacy has enabled a two-way interaction between students and teachers, which greatly enhances the learning experience, as compared to the one-dimensional nature of previous ways (Hague & Payton, 2011).

From an IPO model perspective, Digital literacy in education is more than being able to find information. It involves students being able to critically evaluate and make judgements on information, and if possible, improve upon that information. Students are able to evaluate the trustworthiness and accuracy of information found on the Internet. Thus, not only does it teach students on how to improve their thinking skills, but also help improve the work of others, and thus, benefiting everyone in society. It can be argued that digital literacy is even more important and needs to be more prevalent in Higher Education, whereby it is at this point where students would often need to equip themselves with a high level of digital literacies and competences before going out into the ever-changing and dynamic labour market that is slowly utilising the latest technology (Santos & Serpa, 2017; Shopova, 2014). In particular, the teaching and learning of Science curriculum is cited as one of the curriculums where digital literacy needs to be emphasised on (Ng, 2012).

Online Shopping and Corporate Settings

Retail outlets are under threat from a new method of shopping; online shopping. There are two ways for businesses to commercialise their products, via retail outlets, which requires a physical shop for one-on-one interaction, and online shopping, in which products are sold online on websites, which is commonly known as E-Commerce (Turban, King, Lee & Viehland, 2002). The difference is that unlike retail outlets, online shopping normally requires a degree of digital literacy for transactions to occur between buyer and seller. Due to retail shopping requiring physical stores, they are limited to the local area in which they are operated in, potentially limiting the number of consumers. With online shopping, products can be sold to consumers regardless of their geographical location. Digital literacy allows online seller to design and customise their own website interface. Consumers, at a push of a button at home, could browse for the goods on the Internet and make purchases, and businesses do not need to waste precious resources on setting up shops. Thus, with digital literacy, transaction is made faster and more cost-saving than ever before. The rise of online shopping only serves to highlight the increasing importance of digital literacy in shopping context (Scott, 2001).

In corporate (enterprise) multinational settings, communication has always been an issue due to the relatively large size of the organisation. This diseconomy of scale is made worse with the difference in language in foreign branches (Canback, Samouel, & Price, 2006). However, this barrier to communication can be overcome by using translation software. These advanced software are capable of translating foreign languages for everyone to understand while giving employees the platform to learn and speak new languages, further reducing the barrier to communication. However, these software require organisations to be digitally literate, and it is vital for organisations to ensure that their own employees are digitally literate by providing the necessary education and training. In terms of management, the capability of software to send messages swiftly and to convene low-cost video conferences has enabled organisations to easily keep track of their branches remotely.

Task-wise with the application of an IPO model, digital literacy has made the production and marketing of products easier. After inputting raw materials and components, products can be designed beforehand by using computer-aided design software, and in advertisements, video editing has made it possible to rearrange video shots to create advertisements to be displayed virtually on the Internet. These tasks normally require an insurmountable amount of efforts, but thanks to computers, they can be done in a short amount of time, effectively increasing the production output of goods and services. Suffice it to say, digital literacy results in increased labour productivity among employees. In today's digital society, gone are the days when job candidates are only encouraged to list information technology (IT) skills such as being proficient in PowerPoint on their resume, when such skills are now considered standard.

Healthcare

Digital literacy has always been a crucial part in healthcare for nurses as well as patients. The traditional paper health records have been replaced to give way for electronic means. This is widely known as e-Health (Oh, Rizo, Enkin & Jadad, 2005). With health records being able to be processed at a click of button, it is now more convenient and less time-consuming to look through a patient's health record. This, however, requires nurses being trained in the use of technology. From an IPO standpoint, Digital literacy is needed for nurses to update a patient's health record by inputting crucial patient data such as medical conditions, weight, age, medical test results, and vital signs. Nurses are then able to process this information in order to administer and prescribe medicines as outputs perfectly suited for each patient

depending on sickness and allergies of patients. There is an increasing trend of transforming traditional healthcare to digital health, which refers to the utilisation of technology in order to increase the rate of healthcare delivery (Meskó, Drobni, Béneyi, Gergely & Gyórfy, 2017).

In healthcare, applications are often used in smartphones. In the event of booking for an appointment, digitally literate patients could simply connect to the Internet in their mobile applications to do so. These applications can assist them manage their health by tracking their daily calorie intake, heart rates, intensity of exercises done per week, diabetes level and blood pressure as well as weight loss or gain. Applications are also means for hospitals to raise health and medical awareness by providing online consultation and forums, giving advices on how to stay healthy and stressing the dangers of tobacco and other unhealthy products. Continuous technological developments in healthcare have saved countless lives and improved the quality of life even more. Health-care services need to maximise the training of nurses in digital literacy to enhance patient care.

Government

The centralisation nature of the public sector means that there are often many procedures and ‘red tapes’ which can cause process to take a long time. This involves numerous paperwork which also needs to be approved from the lower level to the higher level. Even with the help of electronic means, these processes can still be time-consuming, albeit faster than before. The embedding of digital literacy onto the public sector workforce is vital for a more faster and efficient public service delivery (Marchionini, Samet, & Brandt, 2003).

Digital literacy is very important as there is an increasing trend towards a digitised government called e-Government (Almarabeh & AbuAli, 2010). While privatisation is rapidly increasing across the globe (Allen, Qian & Qian, 2005), some services continue to serve under government control such as provision of water, police force, national defense and immigration. Common examples of public services include renewal of identification card, passport and driving licence, application for extract from criminal register or commercial register, as well as the forms required for starting a business. All of these processes require public sector employees to have a certain amount of digital literacy. These services are often displayed in government websites, and these websites must be maintained and kept up to date by employees who are digitally literate. Above all, from the viewpoint of an IPO model, digital literacy can help increase the rate of civic participation that enables people to voice their opinions. The technology in hand will allow a faster delivery of messages, whereby the government can give due feedback thanks to a two-way communication in technology which will further revolutionise democratic practices (Robertson & Vatrappu, 2010).

Hindrance of Digital Divide

For this subheading, the paper would be looking at the negative social and economic consequences brought about by digital divide from the perspectives of individuals, firms and governments. Digital divide is referred to as the gap in usage of technology, which could either be demographics or geographic in nature (Chen & Wellman, 2004). Digital divide is a worldwide phenomenon that plagues almost every country. It is more pronounced in countries that adopts a market economy, where the economy is influenced by fluctuation in prices, supply and demand (Moroney & Lovell, 1997), leading to unequal distribution of income and wealth among the populace, which in turn, leads to the classification of rich and poor people.

Although the number of people with access to computers and the Internet continues to soar across the globe, the digital divide also continues to grow at an alarming rate (Chen & Wellman, 2004). It is important to note that some people lack the means and understanding to utilise the technology in hand due to insufficient digital literacy. Despite the best attempts of governments in investing huge sums into technology and Internet infrastructure, the populace has no clue as to how to properly utilise this technology, and the vicious cycle continues in which economic and social progress is halted (Hargittai, 2001).

Economic Implications

One of the biggest issues associated with digital divide is the uneven access to technology. This is due to the costs associated with implementing a whole technological infrastructure. Countries with a healthy gross-domestic product each year would be able to slowly invest in this infrastructure for public consumption. However, for those without the means, will suffer the economic consequences of not being able to do so. A country's access to technological advancements is noted to directly correlate with an economy's development (Borensztein, De Gregorio & Lee, 1998), and vice versa.

Unfortunately, even after government spending huge sums of investments in technology, not all people have the technical know-how on how to utilise technology. This brought about the need to educate the populace in digital literacy which increases costs further in addition to costs of investment and implementation of infrastructure. Lack of IT skills would force them to shun the newly-bought technology giving rise to storage costs. Furthermore, with every update to technology, the cost is not getting cheaper. Thus, under-developed countries would be left behind due to a lack of funds to bridge the digital gap. Worse, they had needlessly spent the money on technological resources that are not being used, making them poorer.

Organisation-wise, redundancy happens due to digital divide. Some organisations are forced to dismiss employees who lack IT skills in order to make way for new employees who are digitally literate. This leads to the rise of ethical and trade union issues, due to workers being unfairly dismissed, and this may lead to lawsuit against the organisation. Costs would also increase should organisations decide to hire lawyers.

In households, it is found that Internet speed varies across regions in the world (Chaudhuri, Flamm & Horrigan, 2005). Internet speed is important as it determines how fast information could be received and delivered. Varying Internet speed results in social inequality whereby there are differences in income and social status. This in turn results in a distinct lack of social capital, whereby members in a society are not able to function as one (Hargittai, 2008; Witte & Mannon, 2010).

Social Costs

A country's social costs may rise when the poor began to resent the rich, especially in countries with a relatively high power-distance index (Hofstede, 2011). The desire to reap the same benefits of technology as those who can afford it have made the poor resorting to theft in order to access this technology either by force or otherwise. This can lead to thousands of dollars' worth of damage to personal property, harm to other individuals and criminal punishment for those who are caught stealing. As the poor people are forced to endure with a meagre amount of income, they are forced to resort to violence and crime just to survive.

For students at school, the inability to connect to the Internet has greatly hindered their ability to finish their homework, as they have no means of external support for knowledge. In extreme cases, especially

among poorer families, students do not even have access to computers, preventing them from doing digital exercises. This potentially leads to students receiving lower grades than their richer counterpart.

For individuals who have access but lack the proper skills, they would be the victims of cyber-bullying. Due to lack of digital literacy and knowledge of cyber-bullying, they are prone to cyber-bullying such as threats and intimidation, harassment, false gossip and rumors and being ridiculed on social media. They would have no means of defending themselves and cyber-bullying has more detrimental effects on younger generation, which may lead to attempted suicides (Qi, Wang, & Xu, 2009).

SOLUTIONS AND RECOMMENDATIONS

The next section will discuss in detail the possible solutions and recommendations on how to bridge the demographic and geographical gap that exists due to digital divide.

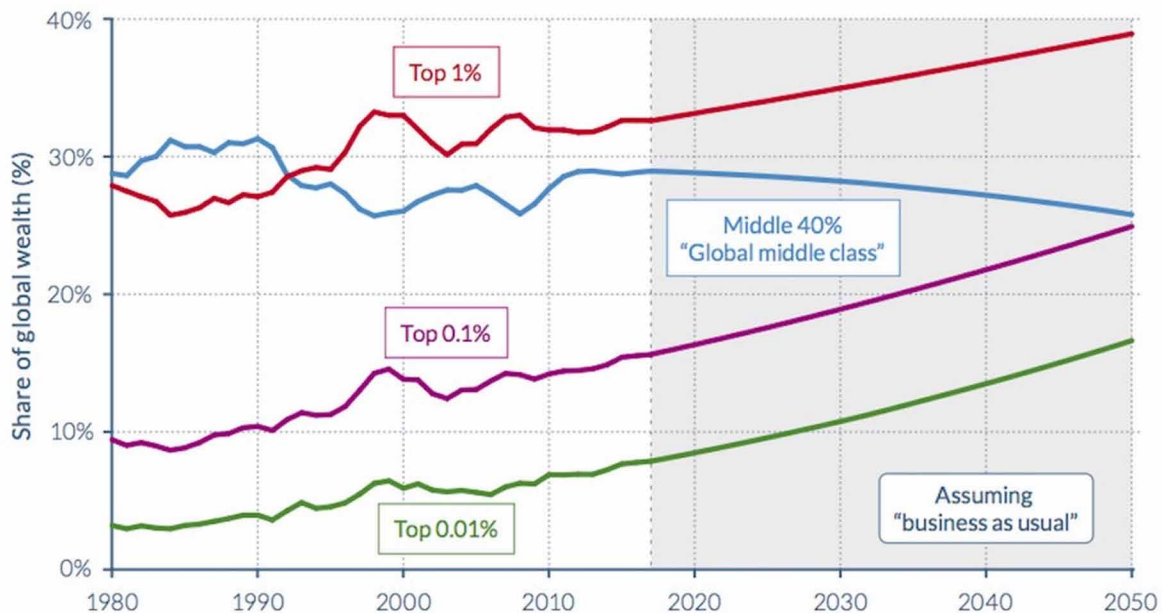
One of the methods to bridge digital divide is by educating the public on the use of technology. Digital divide is also caused by the population lacking in digital literacy despite having access to technology. Thus, to ensure that government investments in technology does not go to waste, it is imperative that the populace across the globe be educated on the use of technology. Although IT classes are the norm, it must be made compulsory in all curriculum from preschool to university level. This is important as it also fosters an atmosphere of technological acceptance. This improved rate of digital literacy will ensure that the technological resources that are readily available are used most efficiently to produce output. In order to encourage public acceptance of technology in general, this paper recommend the application of the widely-used Technology Acceptance Model (Chuttur, 2009) Based on this model, the public will be more accepting towards technology should they perceive that the technology in question is useful and easy-to-use.

In order to expose children in poor families to digital literacy, it is recommended that loans for electronic devices are to be offered. The purpose of this is to expose the populace from an early age to learning technology. Another alternative to loans is the offering of affordable laptops to children, akin to the one laptop per child initiative (Kraemer, Dedrick & Sharma, 2009). This is done by selling low-cost laptops that have the basic features that a child needs to learn digital technology. Unfortunately, not all countries can provide Internet access to its population due to the cost requirement. In this case, this paper recommends that instead of a whole area, only the poorest of remote areas be given relatively cheap Internet access. One such initiative was done by Free Basics (Sen, Pirzada, Phokeer, Farooq, Sengupta, Choffnes & Gummadi, 2016, November), brought about by a synthesis of corporations, albeit this time all services must be made available. It is urged that the private sector does a similar initiative in order to bridge the information gap in poor areas. In addition, to prevent the misuse of any information, the implementation of digital legislations would serve as rules and regulations for children to adhere to regarding the use of information (Qi, Wang, & Xu, 2009), which can potentially help reduce cybercrime.

For countries where there are resistances to adopting new technologies, fostering the right atmosphere is crucial. When possible, awareness campaigns must be initiated to influence a change in mindset. The public must be exposed on the importance of digital literacy especially among the senior citizens. Albeit this method sounds simple, it is by no means easy to implement as it involves conjuring multiple campaigns which can be costly. Public libraries must also be built in countries where digital divide is high. For people who cannot afford technology, public libraries are a means of last resort as a source of information (Kinney, 2010). The facilities in public libraries can also provide free training courses in order to educate the populace to be digitally literate.

However, the settings of these public libraries may not be enough to reach the populace, which will still result in prevailing digital divide. Thus, an initiative must be taken to ensure that automation is implemented in libraries (Mutula, 2005). With an automated system in libraries, not only the usual procedures of library cataloging and lending be more efficient, but ensure faster transition of information to the public, bridging digital divide.

Figure 2. Unequal Distribution of Wealth across the Globe (Source: World Inequality Report, 2018)



Finally, this article would like to implore wealthy organisations to devote themselves to charitable causes. Figure 2 shows that the richest 1% is twice as wealthy as the poorest 50%. As mentioned above, the nature of capitalism means that income and wealth will always be distributed unequally. It is forecasted that this inequality gap will worsen and widen, and if no action is taken, the poor will never be able to break their vicious cycle of poverty, leading to increased crime rate. The unequal wealth distribution means that the rich will always have more than adequate wealth to survive while the poor are scraping by with just a meagre amount of wealth.

Calls for multibillion dollar organisations to submit to charity must be initiated. The giving pledge is a prime example of a campaign to encourage such organisations to contribute a majority of their wealth for philanthropic causes (Chuang, 2015). The kindness to philanthropic causes will not go unrewarded especially if news of this spread throughout the Internet, as image and reputation will be enhanced greatly, to the point where there will be a build-up of followers and consumers, potentially increasing long-term sales (Porter & Kramer, 2002). These organisations are often seen as a bastion of selflessness and role models, potentially causing other organisations to follow suit.

FUTURE RESEARCH DIRECTIONS

Despite the topic revolving digital literacy is still at its infancy, one can notice an emerging trend toward digital literacy as more literature emerges. This highlights the potential of this topic, and this article paves the way for future researchers to delve deep into the importance of digital literacy in more various sectors. More studies needed on the application of the IPO model as the IPO model can be applied to many different scenarios that involves input, process and output. The study of other types of literacy that are similar to digital literacy such as information literacy, media literacy and computer literacy could also be pursued by future researchers. Future researchers could also add on and provide solutions and recommendations from more different perspectives on how to solve digital divide. Above all, a research into unequal wealth distribution that is causing global digital divide would be very useful.

CONCLUSION

In conclusion, the article has provided a definition for digital literacy, in the application of an IPO model, although the concept itself remains unchanged from other definitions of digital literacy in literature. The article also provided a brief overview of having digital literacy in different sectors or industries. e-Learning, e-Commerce, e-Health and e-Government are rising in importance, highlighting the notion that the population needs to have a degree of digital literacy. For a country to achieve economic and social prosperity, its population need to be digitally literate to utilise its technological resources, as humanity is currently living in the digital society, where people are heavily-reliant on digital technology. A lack of digital literacy, as well as the uneven access to technology is the cause of digital divide. Digital divide has detrimental effects, especially to an underdeveloped country, and everyone from governments and organisations down to individuals play an important role in solving digital divide. The solutions and recommendations provided above serve as a general guideline on how to solve the issue of digital divide.

REFERENCES

- Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., . . . Bengio, Y. (2016). *Theano: A Python framework for fast computation of mathematical expressions*. arXiv preprint arXiv:1605.02688
- Ali, A. H. (2011). The power of social media in developing nations: New tools for closing the global digital divide and beyond. *Harv. Hum. Rts. J.*, 24, 185.
- Allen, F., Qian, J., & Qian, M. (2005). Law, finance, and economic growth in China. *Journal of Financial Economics*, 77(1), 57–116. doi:10.1016/j.jfineco.2004.06.010
- Almarabeh, T., & AbuAli, A. (2010). A general framework for e-government: Definition maturity challenges, opportunities, and success. *European Journal of Scientific Research*, 39(1), 29–42.
- Bawden, D. (2001). Information and digital literacies: A review of concepts. *The Journal of Documentation*, 57(2), 218–259. doi:10.1108/EUM0000000007083
- Bawden, D. (2008). Origins and concepts of digital literacy. *Digital literacies: Concepts, policies and practices*, 30, 17-32.

- Berners-Lee, T., Cailliau, R., Groff, J. F., & Pollermann, B. (1992). World-wide web: The information universe. *Internet Research*, 2(1), 52–58. doi:10.1108/eb047254
- Borensztein, E., De Gregorio, J., & Lee, J. W. (1998). How does foreign direct investment affect economic growth? 1. *Journal of International Economics*, 45(1), 115–135. doi:10.1016/S0022-1996(97)00033-0
- Bruce, C. S. (1999). Workplace experiences of information literacy. *International Journal of Information Management*, 19(1), 33–47. doi:10.1016/S0268-4012(98)00045-0
- Buckingham, D. (2008). What do young people need to know about digital media. *Digital literacies: Concepts, policies and practices*, 73–88.
- Buckingham, D. (2010). Defining digital literacy. In *Medienbildung in neuen Kulturräumen* (pp. 59–71). VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92133-4_4
- Buckingham, D. (2015). Defining digital literacy-What do young people need to know about digital media? *Nordic Journal of Digital Literacy*, 10, 21–35.
- Bushnell, D. S. (1990). Input, process, output: A model for evaluating training. *Training and Development Journal*, 44(3), 41–44.
- Canback, S., Samouel, P., & Price, D. (2006). Do diseconomies of scale impact firm size and performance? A theoretical and empirical overview. *ICFAI Journal of Managerial Economics*, 4(1), 27–70.
- Capurro, R. (2009). Digital ethics. In *Global Forum Civilization and Peace. The Academy of Korean Studies and Korean National Academy*. UNESCO.
- Chaudhuri, A., Flamm, K. S., & Horrigan, J. (2005). An analysis of the determinants of internet access. *Telecommunications Policy*, 29(9–10), 731–755. doi:10.1016/j.telpol.2005.07.001
- Chen, W., & Wellman, B. (2004). The global digital divide—within and between countries. *ITandSociety*, 1(7), 39–45.
- Chuang, J. A. (2015). Giving as Governance: Philanthrocapitalism and Modern-Day Slavery Abolitionism. *UCLA Law Review. University of California, Los Angeles. School of Law*, 62, 1516.
- Chuttur, M. Y. (2009). Overview of the technology acceptance model: Origins, developments and future directions. *Working Papers on Information Systems*, 9(37), 9–37.
- Cruz-Jesus, F., Oliveira, T., & Bacao, F. (2012). Digital divide across the European Union. *Information & Management*, 49(6), 278–291. doi:10.1016/j.im.2012.09.003
- Curry, A., Flett, P., & Hollingsworth, I. (2006). *Managing Information & Systems: The Business Perspective*. Routledge.
- Driscoll, M. (2002). Blended learning: Let's get beyond the hype. *E-learning*, 1(4), 1–4.
- Eshet, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia*, 13(1), 93–106.
- Gilster, P., & Glister, P. (1997). *Digital literacy*. Wiley Computer Pub.
- Goel, A. (2010). *Computer fundamentals*. Pearson Education India.

- Grady, J. O. (1995). *System Engineering Planning and Enterprise Identity* (Vol. 7). CRC Press.
- Hague, C., & Payton, S. (2011). Digital literacy across the curriculum. *Curriculum Leadership*, 9(10).
- Hargittai, E. (2001). *Second-level digital divide: Mapping differences in people's online skills*. arXiv preprint cs/0109068
- Hargittai, E., & Hinnant, A. (2008). Digital inequality: Differences in young adults' use of the Internet. *Communication Research*, 35(5), 602–621. doi:10.1177/0093650208321782
- Helbig, N., Gil-García, J. R., & Ferro, E. (2009). Understanding the complexity of electronic government: Implications from the digital divide literature. *Government Information Quarterly*, 26(1), 89–97. doi:10.1016/j.giq.2008.05.004
- Hoffman, D. L., & Novak, T. P. (1998). *Bridging the racial divide on the Internet*. Academic Press.
- Hoffman, M., & Blake, J. (2003). Computer literacy: Today and tomorrow. *Journal of Computing Sciences in Colleges*, 18(5), 221–233.
- Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture*, 2(1), 8. doi:10.9707/2307-0919.1014
- Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, 56(1), 517–543. doi:10.1146/annurev.psych.56.091103.070250 PMID:15709945
- Internet World Stats. (2018, Dec 31). *Internet Users in the World by Region*. Retrieved from <https://www.internetworldstats.com/stats.htm>
- Jaeger, P. T., Bertot, J. C., Thompson, K. M., Katz, S. M., & DeCoster, E. J. (2012). The intersection of public policy and public access: Digital divides, digital literacy, digital inclusion, and public libraries. *Public Library Quarterly*, 31(1), 1–20. doi:10.1080/01616846.2012.654728
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2013). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1), D199–D205. doi:10.1093/nar/gkt1076 PMID:24214961
- Kinney, B. (2010). The Internet, public libraries, and the digital divide. *Public Library Quarterly*, 29(2), 104–161. doi:10.1080/01616841003779718
- Kipnis, D. (2012). *Technology and power*. Springer Science & Business Media.
- Koltay, T. (2011). The media and the literacies: Media literacy, information literacy, digital literacy. *Media Culture & Society*, 33(2), 211–221. doi:10.1177/0163443710393382
- Kraemer, K. L., Dedrick, J., & Sharma, P. (2009). One laptop per child: Vision vs. reality. *Communications of the ACM*, 52(6), 66–73. doi:10.1145/1516046.1516063
- Livingstone, S. (2004). Media literacy and the challenge of new information and communication technologies. *Communication Review*, 7(1), 3–14. doi:10.1080/10714420490280152
- Lopez, M. H., Gonzalez-Barrera, A., & Patten, E. (2013). *Closing the digital divide: Latinos and technology adoption*. Academic Press.

- Marchionini, G., Samet, H., & Brandt, L. (2003). Digital government. *Communications of the ACM*, 46(1), 25–27.
- Martin, A. (2005). DigEuLit—a European framework for digital literacy: a progress report. *Journal of eLiteracy*, 2(2), 130–136.
- Martin, A. (2008). Digital literacy and the digital society. *Digital literacies: Concepts, policies and practices*, 30, 151–176.
- Meskó, B., Drobni, Z., Bényei, É., Gergely, B., & Györfy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. *mHealth*, 3, 3. doi:10.21037/mhealth.2017.08.07 PMID:29184890
- Minghetti, V., & Buhalis, D. (2010). Digital divide in tourism. *Journal of Travel Research*, 49(3), 267–281. doi:10.1177/0047287509346843
- Moroney, J. R., & Lovell, C. A. K. (1997). The relative efficiencies of market and planned economies. *Southern Economic Journal*, 63(4), 1084–1093. doi:10.2307/1061243
- Mutula, S. M. (2005). Bridging the digital divide through e-governance: A proposal for Africa's libraries and information centres. *The Electronic Library*, 23(5), 591–602. doi:10.1108/02640470510631308
- Ng, W. (2012). Why digital literacy is important for science teaching and learning. *Curriculum and Leadership Journal*, 10(10).
- Oh, H., Rizo, C., Enkin, M., & Jadad, A. (2005). What is eHealth (3): A systematic review of published definitions. *Journal of Medical Internet Research*, 7(1). Advance online publication. doi:10.2196/jmir.7.1.e1 PMID:15829471
- Pool, C. R. (1997). A new digital literacy a conversation with Paul Gilster. *Educational Leadership*, 55, 6–11.
- Porter, M. E., & Kramer, M. R. (2002). The competitive advantage of corporate philanthropy. *Harvard Business Review*, 80(12), 56–68. PMID:12510538
- Qi, M., Wang, Y., & Xu, R. (2009). Fighting cybercrime: Legislation in China. *International Journal of Electronic Security and Digital Forensics*, 2(2), 219–227. doi:10.1504/IJESDF.2009.024905
- Rabitz, H., Aliş, Ö. F., Shorter, J., & Shim, K. (1999). Efficient input—Output model representations. *Computer Physics Communications*, 117(1-2), 11–20. doi:10.1016/S0010-4655(98)00152-0
- Report, W. I. (2018). *Presentation of the World Inequality Report 2018*. Retrieved from: <https://wir2018.wid.world/files/download/wir-presentation.pdf>
- Rhodes, J. A., & Robnolt, V. J. (2009). Digital literacies in the classroom. *Handbook of adolescent literacy research*, 153–169.
- Robertson, S. P., & Vatrappu, R. K. (2010). Digital government. *Annual Review of Information Science & Technology*, 44(1), 317–364. doi:10.1002/aris.2010.1440440115
- Robinson, L., & Bawden, D. (2001, June). Libraries and open society; Popper, Soros and digital information. *Aslib Proceedings*, 53(5), 167–178. doi:10.1108/EUM00000000007051

- Santos, A. I., & Serpa, S. (2017). The importance of promoting digital literacy in higher education. *International Journal of Social Science Studies*, 5(6), 90. doi:10.11114/ijsss.v5i6.2330
- Scott, A. (2001). Online shopping on the rise. *Internal Auditor*, 58(1), 15–15.
- Sen, R., Pirzada, H. A., Phokeer, A., Farooq, Z. A., Sengupta, S., Choffnes, D., & Gummadi, K. P. (2016, November). On the Free Bridge Across the Digital Divide: Assessing the Quality of Facebook's Free Basics Service. In *Proceedings of the 2016 Internet Measurement Conference* (pp. 127-133). ACM. 10.1145/2987443.2987485
- Shopova, T. (2014). Digital literacy of students and its improvement at the university. *Journal on Efficiency and Responsibility in Education and Science*, 7(2), 26–32. doi:10.7160/eriesj.2014.070201
- Sidney Howland, J. (1998). The 'digital divide': Are we becoming a world of technological 'haves' and 'have-nots?'. *The Electronic Library*, 16(5), 287–289. doi:10.1108/eb045651
- Turban, E., King, D., Lee, J., & Viehland, D. (2002). *Electronic commerce: A managerial perspective 2002*. Prentice Hall.
- Warschauer, M. (2002). Reconceptualizing the digital divide. *First Monday*, 7(7). Advance online publication. doi:10.5210/fm.v7i7.967
- Williams, P. (2006). Exploring the challenges of developing digital literacy in the context of special educational needs communities. *Innovation in Teaching and Learning in Information and Computer Sciences*, 5(1), 1–16. doi:10.11120/ital.2006.05010006
- Witte, J. C., & Mannon, S. E. (2010). *The Internet and social inequalities*. Routledge. doi:10.4324/9780203861639
- Zelle, J. M. (2004). *Python programming: an introduction to computer science*. Beedle & Associates, Inc.

ADDITIONAL READING

- Bélanger, F., & Carter, L. (2009). The impact of the digital divide on e-government use. *Communications of the ACM*, 52(4), 132–135. doi:10.1145/1498765.1498801
- Davies, J. (2008). Pay and display: the digital literacies of online shoppers. *Digital Literacies. Concepts, Policies and Practices*, 227-248.
- Efimova, L., & Grudin, J. (2008). Crossing boundaries: Digital literacy in enterprises. *Digital literacies: Concepts, policies, practices*, 203-226.
- Knobel, M. (2008). *Digital literacies: Concepts, policies and practices* (Vol. 30). Peter Lang.
- Knobel, M., & Lankshear, C. (2008). Digital literacy and participation in online social networking spaces. *Digital literacies: Concepts, policies and practices*, 249-278.
- Ng, W. (2012). Can we teach digital natives digital literacy? *Computers & Education*, 59(3), 1065–1078. doi:10.1016/j.compedu.2012.04.016

Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge University Press. doi:10.1017/CBO9781139164887

Taylor, R. W., Fritsch, E. J., & Liederbach, J. (2014). *Digital crime and digital terrorism*. Prentice Hall Press.

KEY TERMS AND DEFINITIONS

Digital Divide: The gap between those who have access to technology and those without. Also include those who have access to technology but are unable to utilise them due to lack of digital literacy.

Digital Literacy: The skills and competences of people in using technological platforms to produce and evaluate information through electronic means.

Digital Society: A society in which everything runs on digital technology where paperless and electronic means are the norm.

Digital Technology: The use of electronic tools and devices to produce and generate data.

E-Government: A form of digitised government where public services are delivered and processed via electronic or digital means.

Healthcare: The provision of medical care and services to the public.

Social Capital: The relationship between members of a society who work together to function effectively for the benefit of the society.

Gender Differences in Access to and Use of ICTs in Nigeria

13

Immanuel Ovemeso Umukoro

 <https://orcid.org/0000-0002-2054-5070>

Lagos Business School, Pan-Atlantic University, Nigeria

Aanuoluwapo Oluwaseun Omolade-Lawal

University College Hospital, Ibadan, Nigeria

Samuel Oyelami Babalola

UAC Nigeria Plc, Nigeria

Kolawole Sunday Akinsumbo

LucrativeTech, Nigeria

Rashida Mebude Aligwa

Tek Experts Limited, Nigeria

Balikis Animasaun Abdul-Jeleel

International Institute of Tropical Agriculture (IITA), Nigeria

INTRODUCTION

There are varying perceptions regarding gender equity in access to and use of information and communication technologies. Some argue that women have been maligned and are disadvantaged compared to their male counterparts while others are of the view that equal opportunities abound for both genders. Sanda and Kurfi (2013) submit that in their quest for empowerment and societal recognition, women have adopted and accepted ICTs as instruments for their advancement and empowerment. Although current information is fragmented, there is evidence suggesting that there are two schools of thought on gender differences in access to, and use of ICTs. The first school of thought argues that there are apparent gender differences in access to and use of ICTs (Kose, 2019; Olatokun, 2009; Nsibirano, 2009; Nsibirano, 2006; Agbonlahor, 2005; Brous, 2005; Stewart, 2004; Hafkin & Taggart, 2001). The second school of thought believes that women have come to a position of equal opportunities and platforms with their men counterparts in access to and use of ICTs (Olsson, Samuelsson & Viscovi, 2019; Bhandari, 2019). This view is also shared by Morahan-Martin (2000) who posit that females have reached parity with males in use of the internet and claims that digital divide in access to and use of ICT has disappeared. Gillwald, Milek and Stork (2010) also submit that women in some parts of the world have access to and make more use of ICTs when compared to men. This second school of thought also believed that male domination of the ICT field is not the case anymore.

When looking at the younger generation, especially the digital natives, the gender difference appears to be obliterated with both boys and girls seeming to be involved and interested in the use of various ICT tools, both at school and elsewhere (CERI, 2010). Furthermore, Tomte (2011) notes that there is

DOI: 10.4018/978-1-7998-3479-3.ch117

the emergence of social media with seemingly more female participation. In light of this, there is a need to explore how the gender issue appears in youngsters' use of and attitudes towards ICTs, and how they perform and interact as producers and consumers of digital contents. The need to bridge the digital divide between men and women in access to and use of information and communication technologies is necessary given that ICTs play essential roles in human development as enablers of meaningful and sustainable development (Gillwald et al. 2010). Furthermore, the United Nations Division for the Advancement of Women (2003) affirms that there is evidence suggesting that access to ICTs can break down the isolation of individuals, enhance their chances of economic and social inclusion as well as provide diverse avenues for women's social, political, and economic empowerment. To this end, ICTs are potential tools for job creation, social integration, women empowerment, poverty alleviation, and information dissemination. Consequently, access to and use of ICTs cannot be gender-defined but rather transcend gender, income, educational level, geographical location as well as social stratification.

Sorenson (2002) reports that widespread access to and use of ICTs do not necessarily give women equal position with men. Although studies have attempted to understand the gender gap, these studies only tell if women are included or excluded from access to and use of ICT without significant consideration of the factors responsible for this divide. Where attempts have been made to understand the causes of the gender divide, these tend to use dissimilar populations such as urban versus rural (Olatokun, 2009), educated versus uneducated, employed versus unemployed, and such likes without attempting a parallel assessment of gender difference using equal or similar criteria such as salary, employment, education level, social status etc. Given the importance of ICTs to women development, Gillwald et al. (2010) focused on ICTs in the context of Women in Development (WID), Women and Development (WAD) and Gender and Development (GAD). Other scholars also argue that studies on women and technology are sparse and often neglected (Kuschel Labra & Díaz, 2018; Orser & Riding, 2018; Dy et al., 2017; Kuschel & Lepeley, 2016). As interests deepen in studies focusing on gender inequality in access to and use of ICTs, there is an indication that the widespread adoption and application of ICTs to everyday life – both at home and the workplace is causing a shift towards an equilibrium in terms of male and female access to and use of ICTs. This study, therefore, assessed the gender difference in access and use of ICTs.

The following questions guided the study.

- i. Do males and females have equal access to ICTs in Nigeria?
- ii. What are ICTs mostly used by both males and females in Nigeria?
- iii. What are the factors that facilitate access to and use of ICTs among males and females in Nigeria?
- iv. What are the constraints to access and use of ICTs among male and female Nigerians?

Statement of Hypotheses

Based on the literature review, we hypothesise that

- i. There is no significant difference in access to ICTs between males and females in Nigeria
- ii. There is no significant difference in the use of ICTs between males and females in Nigeria
- iii. There is no significant difference between factors that facilitate access to and use of ICTs among males and females in Nigeria.
- iv. There is no significant difference between factors constraining access and use of ICTs among males and females in Nigeria.

BACKGROUND

13

Information and Communication Technologies – An Overview

Information and communication technology is a collective term for all technologies that facilitate the processing, access to, storage, manipulation, use and exchange of information between a sender and receiver of information across various media. ICT refers to associated technologies that facilitate all activities in the information life cycle. The importance of ICTs cuts across all aspects of human endeavours as almost every industry – agriculture, commerce, education, health, hospitality and tourism, manufacturing, and entertainment is driven by ICT. Consequently, Vernon (2001) viewed ICT as a collective word covering all those technologies, both hardware and software dedicated to the capture, storage, and processing, transmission, and presentation of information. Beyond this definition, the application of ICTs in driving socio-economic activities has had a significant impact on people's quality of life. This is because ICTs are applied in health delivery, engineering, business, and agriculture, military, security, law, politics and governance, all aspects of arts, science and education, among others.

Similarly, Tiarniyu (2003) conceived ICTs as electronic technologies for creating, acquiring, storing, processing, communicating and using information. This goes to show that ICTs are powerful tools for acquiring information that can aid better decision making. The use of ICTs involves the process in which computers and other related machines are used in the dissemination and retrieval of information, a technological approach which is now an integral part of human existence and has become a tool to access, receive, store, retrieve, and disseminate information. Information and communication technologies provide a platform for rapid global mobilisation and dissemination of information among people, industries, businesses and professions with the awareness of its benefits putting one at an edge over those who are unaware. This is one primary reason why there is a growing convergence of businesses, professionals, professions and industries through the leverage of ICTs for strategic advantage and global competitiveness.

The adoption and application of ICTs to work processes and human activities, in general, have been massive since the advent of the Internet. Dominick (2002) opines that the Internet brings the cost of communication to a level where many can afford it and thus making content creation and publication easy and faster. With the possibility of accessing an audience of millions of Internet users, creating a whole new class of mass communicators has been enabled by the use of ICTs. The role of ICTs in the production and use of knowledge has also become increasingly important as the basis for what is often coined as 'the new knowledge economy', which can be seen as a response to increasing global competition between countries and regions drawing on knowledge as the main asset. To ensure that society and individuals are included in this development, the idea of 'digital inclusion' has been advocated for in order to enhance the social, political and economic inclusion of all socially marginalised groups in which gender has been embedded. The use of ICTs in everyday activities has become a necessity for all as evident in education, commerce, health, agriculture, technology, religion, government, politics, entertainment, tourism, and the likes. This is due to increasing awareness of the use of ICTs for strategic advantage, a convergence of people, professionals and professions, media and formats, computing and office technologies.

In emphasising the importance of information and communication technologies to human endeavours, Nedum (2006) reports that ICTs have impacted education in the way teaching and learning are now delivered. This delivery mechanism takes the form of motion picture or film, film strip, slide projection, overhead transparency, Teletext and video text, reprography and photography, scanning and

faxing microchips, microfilms and microfiche, voice mail, voice processing, use of computers and the internet, and multimedia systems. These are all essentials of computer-based training (CBT), e-learning, and computer-mediated learning that help in interactive processing information services. Similarly, Sanda and Kurfi (2013) posit that ICTs are wielding an unimaginable influence on global activities as they have restructured politics and governance, economies, cultures, health delivery and administrative systems, and the likes across many countries and regions around the world. These ICT-driven changes are creating several development opportunities by contributing to information dissemination, providing a plethora of communication capabilities, increased access to technology and knowledge among others with individuals and groups having embraced ICTs globally as a tool for enhancing their varied interests. The importance of ICTs to human endeavours cannot be overemphasized as their application cuts across all human endeavours. Consequently, there is heightened emphasis on the use of ICTs for global development, leading to many innovations resulting from exposure and access to ICTs that provide people access to information beyond time and border.

Gender Issues in Access to and Use of ICTs

According to Asiabaka (2010), access refers to the ways and means in which individuals, communities and institutions are exposed to ICT and takes into consideration such elements as affordability, availability of the technologies, access point, and the times at which the technologies are available. The societal stand regarding gender disparity in access to and use of ICTs is that men have been the dominant users when compared to women. In light of this, Orser and Riding (2018) note that social influences may be more salient for women compared to men in forming an intention to use technology. In a study that determined the use of e-government in Turkey, Kose (2019) notes that there is a gender gap in use of e-government services as Turkish females are less likely to employ e-government tools compared to males. Tomte (2011) asserts that there has also been a point of departure for policymakers in their efforts to make plans and activities that address gender disparity in access to and use of ICTs. Women have long been aware of the power of knowledge and information sharing as a strategy for mobilising development efforts, economic empowerment and advancement.

The Beijing 1995 world conference on women recommended that women have equal access to economic resources. These include science and technology, vocational training, land, credit, information, communication and markets to further the advancement and empowerment of women and girls. This also includes the enhancement of their capacities to enjoy the benefits of equal access to these resources through international cooperation. Similarly, the conference emphasised the need to increase the participation of women to expression and decision-making in and through the media and information communication technologies as well as to promote a balanced and non-stereotyped portrayal of women (Sanda & Kurfi, 2013). If given the enabling environment, ICTs can be used to empower women through access to and use of information for informed decision making both at homes and for business. Furthermore, women can also use ICTs to harness generated information into new and localised knowledge for improving their businesses, accessing quality healthcare, and also share with others in the community, thus contributing towards their empowerment, self-determination and well-being. Unlike high prevalence in mobile phone use, Gillwald et al. (2010) report that being female and living in rural settlement reduced the probability of knowing what the Internet is, as well as using Internet-enabled ICTs when compared to female living in urban areas with increased access to ICTs. They further stated that similar education, employment and financial background create similar awareness in males and female in terms of access to and use of ICTs. However, women's failure to get equal education, employment as well as financial resources forms

the more significant part of their marginalisation in access to and use of ICTs. Umukoro and Tihamiyu (2017) also report that males have a higher disposition to the adoption and use of e-services delivered via information systems which are majorly an assemblage of ICTs.

Howell and Lundall (2000) emphasise that the effective use of ICT in a country impacts strongly on the competitiveness of the country's economy within the global market place and the ability of governments to deliver on their social goals. However, Heintz (2012) noted that gender-based constraints – including responsibilities for unpaid care and household work, social norms and gender roles, differences in women's access to and control over assets and finance, unequal investments in the capabilities of girls when compared to boys limit women's choices relative to men with regard to employment. In an attempt to explain this supposition, Sanda and Kurfi (2013) define gender as a socio-economic variable for analyzing roles, responsibilities, constraints, and needs of men and women within a given context and also refer to the social and cultural constructs that each society assigns to behaviours, characteristics and values attributed to men and women. The basis of the construct lies underneath the idea that these gender characteristics are natural or intrinsic, and therefore, unalterable. These gender attributes are shaped by ideological, historical, religious, ethnic, economic and cultural determinants which are usually translated into socio-economic and political inequalities, where men's activities and their attributes are perceived as essentially superior to women's. These assumptions are capable of heightening both digital and social exclusion at the disadvantage of women. The consequence could also include women missing out on critical positions in the society that could benefit the larger society.

In Nigeria and Sub-Saharan Africa (SSA), Okunna (2000) opines that gender relations are characterised by much imbalance to women's disadvantage by keeping women in subordinate positions to their men counterparts. Women are often considered to be 'naturally' nurturing – a characteristic linked to their reproductive capacity as child bearers and attribution that has permeated the field of science and technology globally. In the developing world, the male subculture sees women and their aspirations subordinate, resulting in a situation in which the marginalisation, trivialisation and the stereotyping of women are becoming glaring aspects of the human society. In consonance, Osabuohien and Karakara (2018) and Agbonlahor (2005) opine that there is a growing notion that women are not given the equal opportunity as men in access to and use of ICTs as males still have a good advantage over women. Some researcher (Madanda & Kabonesa, 2007; Nsibirano, 2006) also corroborate that women constitute a minority of users of the Internet and other related technology. Similarly, Manago and Pacheco (2019) note that young adult males and those who have been to high school were found to be more likely to use ICTs such as the internet. While this is true, there are also indications that much of society has changed over the years relating to female participation in leadership roles, politics, religion and the larger business environment. However, the pace of change in female inclusion in access to and use of technologies has been slow in developing and emerging economies when compared to developed countries. Although some scholars (Bhandari, 2019; Olsson et al., 2019; Morahan-Martin, 2000) posit that females have reached parity with males in use of the internet and related ICTs leading to a gradual disappearance of the gendered digital divide in access to and use of ICT, scholarly opinions are still nuanced in respect to the gender difference in access to and use of ICTs.

Constraints to Access and Use of ICTs: A Look Beyond Gender

With growing literature on women marginalisation in access to and use of ICTs, studies have pointed two opposite directions regarding the challenges to access and use of ICTs. Some studies (Olatokun, 2009; Heintz, 2010) show that government and society have not done enough to encourage women's effort

toward inclusive access to and use of ICTs. On the other hand, others (Sanda & Kurfi, 2013; Morahan-Martin, 2000) have shown that same factors such as low level of education, poor skill set, inadequate technological infrastructure, social and economic problems, rapid technological advancement, etc affect both male and female access to and use of ICTs. While men seek to fully integrate themselves into the technology industry thereby optimising access to and use of ICTs, women, on the other hand, lack the necessary skills to make use of ICTs and have many negative attitudes about them. Rather than address these challenges, they are often lackadaisical – so much depending on their male counterparts to get ICT related issues sorted for them. The African Gender Institute (2003), however, reports that certain aspects of culture constrain women's access to and use of ICTs especially for married women as some families still fear that women's access to ICTs may bring them into parity with their husband and erase the cultural distance upon which respect and other norms are built upon. Notwithstanding, there are some changes in women's use of ICTs in recent years as they have begun developing skills needed to operate this new technology effectively (Sanda & Kurfi, 2013) to support their businesses and aid their communication.

In recognition of women's critical role in the development of less industrialised societies, Scanlan (2004) posits that it is undoubtedly a fact that regardless of economic development level, population pressure, democratisation, globalisation, or region, women are at the forefront of development. Consequently, they must furnish themselves with requisite skills such as education, entrepreneurship, political participation, self-esteem, etc that men have explored to be placed ahead of their women counterparts. The Nigerian woman is characterised by low self-esteem because society has continued to regard her as unimportant and inferior to her male counterpart (Sanda & Kurfi, 2013; Okunna, 2000). Similarly, Siddiq and Scherer (2019) posit that prior research suggests that boys consider their ICT self-efficacy to be higher than that of girls. However, their meta-analysis study of whether the same pattern holds for students' actual performance on ICT literacy tasks showed that the gender differences in ICT literacy were significant, positive, and favoured girls though not very severe. Although this narrative is gradually changing with a limited number of women as CEOs and political office holders, a large percentage of women are still on the negative side of the socio-economic, ethno-cultural, and digital divide. Women must, therefore, address the issue of low self-esteem and begin to see themselves capable of doing what men can do in terms of access to and use of information and communication technologies for strategic advantage. Although progress has been witnessed in the area of techno-entrepreneurship where Nigerian women have become prominent in the technology industry, societal challenges have prevented women from taking up challenging tasks. Notwithstanding, David-West and Nwagwu (2018) report that efforts at bridging gender gap (SDG 5) and related use of ICTs has seen the deployment of innovative technologies in Nigeria's financial system that encourage women's access to and use of formal digital financial services. These services are delivered via mobile phones and other digital channels and have helped close the gender gap in use of ICTs and access to financial services.

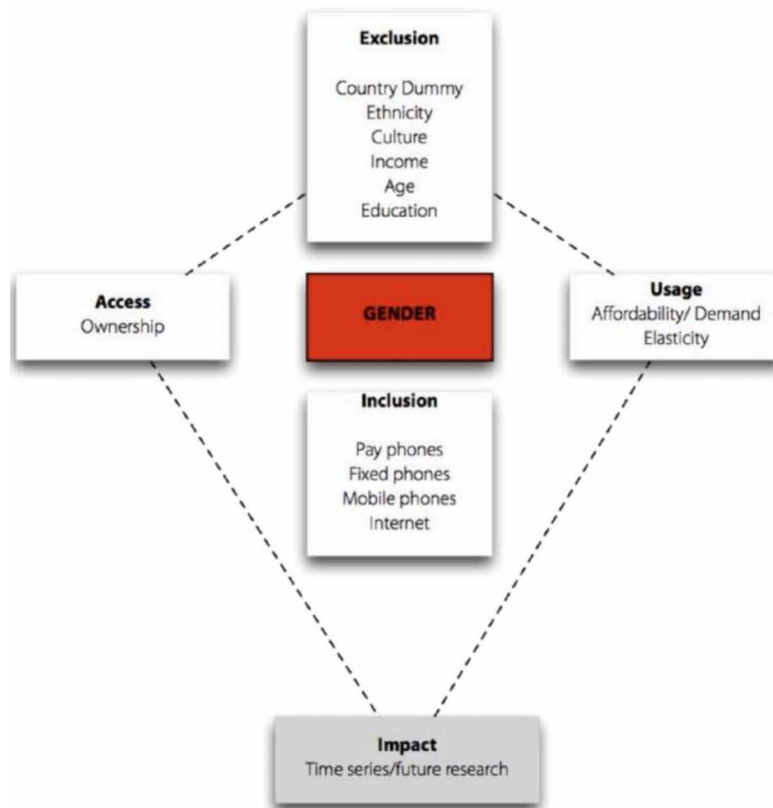
Huyer and Hafkin (2007) point that a range of social, economic, and political factors affect and frame the gender divide, including socio-cultural barriers to technology use, education and skills levels, employment and income patterns, media and content, privacy and security, as well as location/mode of access. According to Gillwald et al. (2010), there is an insignificant gender difference in access to ICTs and their use when men and women share similar backgrounds. They further note that because of unequal access to the enablers of access to and use of ICTs, women generally have restricted access to ICTs and this increases as the technologies and services become more sophisticated and expensive, requiring higher levels of income and education to access and to operate. Respectively, women in lower income groups or rural areas may have more in common with men in similar situations with regards to access and us-

age of ICTs than with women who are more educated or from higher income groups. Also, as mobile phones become increasingly pervasive, their usage becomes limited by the high cost of communications.

Research Model and Conceptual Framework

Different research on the gender difference in access to and use of ICTs have been carried out, yet very few frameworks exist on the subject – notably the capability approach (Olatokun, 2009) and inclusivity model (Gillwald et al., 2010). The inclusivity model guides this study due to its fit for the variable of interest such as ICTs, gender inclusion and exclusion, access and use. The model in Figure 1 shows the relationship between predictors and their interaction with gender. *Inclusion* refers to those strategies or set of activities and facilities aimed at recruiting people into an information system and are further meant to facilitate access to and use of ICTs as well as it relates to gender society (Sorenson, 2002). *Exclusion* refers to measures or factors (collectively referred to as country dummy) that create or determine inequality between the ‘haves’ and ‘have-not’. These factors include culture, age, income, education and ethnicity. *Access* has to do with ownership either permanently or for the period of use (paid for) which has to do with social, economic and cultural factors. *Use* refers to the deployment or utilisation of systems or services consequent of affordability and skills required to manipulate ICTs for strategic advantage efficiently.

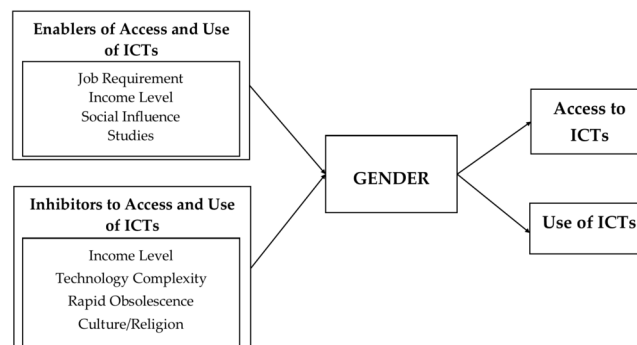
Figure 1. The Inclusivity Model (Gillwald et al., 2010).



Conceptual Framework

For suitability and adequate coverage variable wise, the inclusivity model was adopted, as shown in Figure 2. The framework portrays a relationship among variables determining access to and use of ICTs among males and female. Enablers of access and use include job requirement, income level, societal influence and academic pursuit. These constructs are perceived to drive access and use of ICTs irrespective of gender. Inhibitors to access and use of ICTs include income level, the complexity of technology, and rapid obsolescence of technology as well as religion/culture. These two groups of constructs determine access and use of ICTs. Income earners and those with jobs requiring ICTs are likely to afford technology as non-access, and use is considered as a personal choice and not due to societal exclusion unlike those with low income. The framework is useful in assessing what variables affect males and females' access to and use of ICTs.

Figure 2. Conceptual Framework of Determinants of Gendered Access to and Use of ICTs



Methodology

This study employed a survey design to understand differences in access to and use of ICTs between males and females randomly selected from a mixed population of working and non-working residents of Ibadan, Oyo State, Nigeria. Ibadan is home to Nigeria's premier university and is characterised by high traffic of economic, social and academic activities with high mobile penetration. Data was collected from 100 respondents made up of 50 males and females each using a questionnaire. The instrument used eleven items to measure access to ICTs and nine items for the use of ICTs using Likert scale. A Thurstone equal interval appearing scale was also used to measure enablers and inhibitors of access to ICTs. An independent sample t-test was used to measure the gender difference in access and use of ICTs as well as the enablers and constraints.

Results

The results of the analysis is presented and interpreted under the following sub-sections:

Demographic Characteristics of the Respondents

Table 1 shows that of the 100 respondents sampled for the study, 50% were males, while 50% were females as well. The age distribution followed as 16-25yrs (24%), 26-35yrs (60%) whereas 39yrs or above accounted for 16%. Furthermore, 7% of respondents had up to secondary education while the majority (93%) had more than secondary education. It was also observed that 50% were students, 41% of workers and 9% were unemployed in terms of occupation. However, 45% were on a salary scale of less than \$170 per month, 28% receives more than \$170 per month whereas 27% were without salary (partly because they are unemployed and also studying. The distribution of the respondents across religion showed that 24% were Islamists, whereas 71% were Christians.

Table 1. Respondents' Demographics

Variables	Category	Frequency	Per cent
Gender	Male	50	50.0
	Female	50	50.0
Age	16-25yrs	24	24.0
	26-35yrs	60	60.0
	36yrs & Above	16	16.0
Educational Level	O'Level or Less	7.0	7.0
	Above O'Level	93	93.0
Occupation	Student	50	50.0
	Working	41	41.0
	Unemployed	9	9.0
Salary Range (in Naira)	less than 20,000	25	25.0
	20,000 - 60,000	20	20.0
	above 60,000	28	28.0
	Missing**	27	27.0
Religion	Islamic	24	24.0
	Christianity	71	71.0
	Others	3	3.0
	Missing**	2	2.0

**Income level and religion remain sensitive variables that are often skipped by respondents in Nigeria

RQ 1: Do males and females have equal access to ICTs in Nigeria?

Table 2 shows that both males and females have equal (50%) access to mobile phones. However, males are reported to have higher (31%) consistency in accessing internet facilities than their female counterparts (27%). On the other hand, access to internet modems is slightly different between males (44%) and females (38%) respectively. It was also observed that access to smartphones, PDAs and Laptops was slightly higher among males (47%) when compared to females (42%). More so, access to ICTs provided by schools and offices (in the workplace) was higher with males (45%) against that of females (41%).

Table 2. Access to ICTs across Gender

<i>I have access to</i>	Categories	Male	Female	Total
Personal Phones	Always (%)	46	50	96
	Sometimes (%)	4	0	4
	Never (%)	0	0	0
Internet Facilities	Always (%)	31	27	58
	Sometimes (%)	19	23	42
	Never (%)	0	0	0
Internet Modem	Always (%)	22	19	41
	Sometimes (%)	22	19	41
	Never (%)	6	12	18
Smart Phones, PDAs/Laptops	Always (%)	36	32	68
	Sometimes (%)	11	10	21
	Never (%)	3	8	11
ICTs in my Office/School	Always (%)	24	19	43
	Sometimes (%)	21	22	43
	Never (%)	5	8	13

RQ 2: What ICTs are mostly used by males and females in Nigeria?

Table 3 reveals that the ICTs mostly used by males and females are mobile phones (98%) followed by Laptops and PDAs (92%), desktop computers (91%) and the internet (89%). Other ICTs are seldom patronised for use as shown with as low as (11%) of the respondents involved in the use of these technologies (other than the mobile phone, computers – *desktops and laptops*, PDAs and the Internet).

Table 3. Frequently Used ICTs by Males and Females

<i>I make use of</i>	Response	Male	Female	Total
Mobile Phone	Often (%)	49	49	98
	Rarely (%)	0	0	0
	Never (%)	1	1	2
Computers (Desktop)	Often (%)	46	45	91
	Rarely (%)	1	1	2
	Never (%)	3	4	7
Laptops and PDAs	Often (%)	47	45	92
	Rarely (%)	0	1	1
	Never (%)	3	4	7
Internet	Often (%)	47	42	89
	Rarely (%)	0	1	1
	Never (%)	3	7	10
Others**	Often (%)	8	2	10
	Rarely (%)	0	1	1
	Never (%)	42	47	89

RQ 3: What factors facilitate/necessitate the access to and use of ICTs among males and females?

Table 4 shows the factors which facilitate/necessitate access to and use of ICTs among males and females. The result shows that academic studies (78%) received the highest response as a factor which drives both males and females in accessing and using ICTs. This was distantly followed by job requirement (58%), social influence (55%) and lastly, income level (49%) respectively.

Table 4. Factors that facilitate/necessitate Access and Use of ICTs

Factors	Response	Access and Use
Job Requirement	Yes (%)	58
	No (%)	41
	Undecided (%)	1
Income Level	Yes (%)	49
	No (%)	50
	Undecided (%)	1
Social Influence	Yes (%)	55
	No (%)	43
	Undecided (%)	2
Studies	Yes (%)	78
	No (%)	21
	Undecided (%)	1

RQ 4: What are the constraints to access and use of ICTs among males and females in Nigeria?

As shown in Table 5, only income level and cost of acquisition of information and communication technologies were the significant constraints to access and the use of ICTs. All other identified factors were not key constraints to access and use of ICTs by males and females.

Table 5. Constraints to Access and Use of ICTs

Constraints	Response	Access and Use
Income Level /cost	Yes (%)	70
	No (%)	30
	Undecided (%)	0
Complexity of Technology	Yes (%)	37
	No (%)	61
	Undecided (%)	2
Rapid Obsolescence of Technology	Yes (%)	40
	No (%)	57
	Undecided (%)	3
Culture/Religion	Yes (%)	11
	No (%)	87
	Undecided (%)	2

Test of Hypotheses

The result of the analysis of independent sample t-test for difference of mean for each of the hypothesis formulated is presented below.

H₀₁: *There is no significant difference in access to ICTs between Nigerian males and females:* Although Table 6 shows that there is a difference in the means of males (2.66) and females (2.57) in respect to access to ICTs, Table 7 shows that this difference between males and females' access to ICTs is not statistically significant ($t=-0.538$, $df=98$ and $p=0.296>0.05$; 1-tailed). *The null hypothesis was, therefore, not rejected.*

Table 6. Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Access to ICT	Male	50	2.57	.375	.053
	Female	50	2.66	1.064	.150

Table 7. Independent Samples Test

Access to ICT	Levene's Test for Equality of Variances		t-test for Equality of Means				
	F		T	Df	Sig. (2-tailed)	Mean Diff	Std. Error Diff
Equal variances assumed	1.445	0.232	-.538	98	.592	-.086	.159
Equal variances not assumed			-.538	60.976	.593	-.086	.159

H₀₂: *There is no significant difference in the use of ICTs between Nigerian males and females:* The result presented in Table 8 shows there is a difference in the use of ICTs between males and females (2.62 and 3.84 respectively) with females using ICTs more than males. At Levene's $p<0.05$ (equal variances not assumed), Table 9 shows that the difference between males and females is not statistically significant ($t=-1.258$, $df= 49.141$ and $p=0.107>0.05$; 1-tailed). *The null hypothesis was, therefore, not rejected.*

Table 8. Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
ICT USE	Male	50	2.62	.260	.037
	Female	50	3.84	6.850	.969

Table 9. Independent Samples Test

Use of ICT	Levene's Test for Equality of Variances		t-test for Equality of Means				
	F	Sig.	t	Df	Sig. (2-tailed)	Mean Diff	Std. Error Diff
Equal variances assumed	5.558	0.020	-1.258	98	0.211	-1.220	0.969
Equal variances not assumed			-1.258	49.141	.214*	-1.220	0.969

H₃: *There is no significant difference in factors that enable access to and use of ICTs between males and females Nigeria:* Table 10 shows there is a difference between the means (2.67 and 4.44) of males and females with respect to enablers of access to and use of ICTs. At Levene's $p > 0.05$ (equal variances assumed), Table 11 further shows that difference between males and females in terms of enablers of access to and use of ICT is statistically significant ($t = -2.502$, $df = 81.534$ and $p = 0.007 < 0.05$; 1-tailed). *The null hypothesis was, therefore, rejected.*

Table 10. Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Enablers of access to and use of ICTs	Male	50	2.67	2.628	0.372
	Female	50	4.44	4.264	0.603

Table 11. Independent Samples Test

Enablers of access to and use of ICTs	Levene's Test for Equality of Variances		t-test for Equality of Means				
	F	Sig.	t	Df	Sig. (2-tailed)	Mean Diff	Std. Error Diff
Equal variances assumed	10.638	0.002	-2.502	98	0.014	-1.773	0.708
Equal variances not assumed			-2.502	81.534	0.014	-1.773	0.708

H₄: *There is no significant difference the inhibitors of access to and use of ICTs between males and females in Nigeria:* Table 12 shows that there is a difference between the means (2.30 and 2.59) of Nigerian males and females. With equal variances assumed (Levene's $p > 0.05$), it is also observed from Table 13 that the difference in inhibitors to access to, and use of ICTs between males and females is not statistically significant ($t = -0.276$, $df = 97.796$ and $p = 0.392 > 0.05$; 1-tailed). *The null hypothesis was, therefore, not rejected.*

Table 12. Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Inhibitors of access to and use of ICTs	Male	50	2.30	5.126	0.725
	Female	50	2.59	5.365	0.759

Table 13. Independent Samples Test

Inhibitors of access to and use of ICTs	Levene's Test for Equality of Variances		t-test for Equality of Means				
	F	Sig.	t	Df	Sig. (2-tailed)	Mean Diff	Std. Error Diff
Equal variances assumed	0.191	0.663	-0.276	98	0.783	-0.290	1.049
Equal variances not assumed			-0.276	97.796	0.783	-0.290	1.049

Solutions and Recommendations

This study found that males and females have equal access to and information and communication technologies, although males were found to be more consistent in accessing internet facilities than their female counterparts. This found support in Morahan-Martin (2010) that there is a gradual shift towards equality in males and females' access to and use of ICTs. Sanda and Kurfi (2013) also posit that in recent years, there have been changes in women's access to and use of ICTs as women are now developing skills and capabilities needed to utilise existing and emerging technologies effectively. Today's emergent digital economy also means that daily routines are now facilitated by basic ICTs and would require the acquisition of IT capabilities in order to effectively use them (Umukoro & Tihamiyu, 2017; Sanda & Kurfi, 2013; Gillwald et al., 2010; Nedum, 2006). As noted by the United Nations Division for the Advancement of Women (2003), access to and use of ICTs can foster socio-economic inclusion of individuals, enhance their chances of a better quality of life as well as provide diverse avenues for women's social, political and economic empowerment.

The study also found that mobile phones, computers (desktops and laptops) personal device assistants (PDAs), and the internet have become mainstream ICTs used by both males and females in Nigeria. This has become a way of life given that these technologies facilitate processes and activities in education, health, business (commerce), banking, tourism and entertainment, and the likes. Howell and Lundall (2000) emphasise that the effective use of ICTs in a country can impact strongly on the competitiveness of that economy within the global marketplace and government's ability to deliver on its social goals. This is because ICTs can help all social strata to be more productive, especially those at the bottom of the pyramid (BoP). More so, the study shows that the internet and other related ICTs aid social inclusion, especially among women whom socio-cultural norms might have worked against in developing countries. Through ownership of smartphones and remote access to the Internet, women now have improved access to information for improved decision making that enhances their quality of lives.

On the enablers of access to and use of ICTs, the study showed that educational demand (studies), job requirement, need for social inclusion and societal influence, as well as income level, are significant factors that drive access to and use of ICTs among males and females. This is strongly supported by the findings of Nedum (2006) that ICTs have become an integral part of education, communication, health and other aspects of human endeavour. As such, the need to be connected on the go, work seamlessly and remotely can drive the adoption and use of ICTs among men and women. Umukoro and Tihamiyu (2017) also corroborated this fact that the deployment of ICTs in education is a critical enabler of access to and use of ICTs. On the other hand, the study showed that among the factors identified as inhibitors to access and use of ICTs, only income level received a unanimous agreed response. Technology

sophistication, rapid obsolescence of technology as well as culture and religion were not considered as major constraints to ICT access and use. These findings are contrary to those of Sanda and Kurfi (2013), Huyer and Hafkin (2007), and Morahan-Martin (2000) who argued that rapid obsolescence of technology, as well as culture and religion, affect gender access to and use of ICTs. It is true that income level determines the standard of living and can be inferred that those with the financial capability, irrespective of gender will always access and use ICTs for strategic advantage.

The study also showed that there is no significant difference in access to ICTs between males and females. This is explained by the fact that males and females are becoming increasingly aware of the benefits of information and communication technologies' application to human activities for personal and national development. There are mixed opinions on the concept of gendered access to and use of ICTs with some researchers positing that gender differences still prevail while others are of the school of thought that a significant level of equality is being attained as a result of a growing awareness among both males and females on the numerous benefits of ICTs. The current study found that there is no significant difference in the use of ICTs among Nigerian males and females. In consonance, Sanda and Kurfi (2013) submit that women in their quest for empowerment and societal recognition have adopted ICTs as instruments for their advancement and empowerment. Also, Partey et al. (2018) report that both men and women have been cited as very responsive to technology adoption under different contexts. The 'closing gap' in access to and use of ICTs reported by Morahan-Martin gave credence to this finding. As stated by Dominick (2002), ICTs' adoption and application to everyday human activities is a sure way to reducing the cost of communication to a level where many can afford them thus creating an environment of inclusivity in the social, political, entertainment, tourism, health, education and other human environments which man must constantly relate with.

This study also found a significant difference between factors that facilitate males' access to/use of ICTs and those of females. This contrasts Morahan-Martin (2000) and Sanda and Kurfi (2013) that same factors such as low level of education, poor skill set, inadequate technological infrastructure, social and economic problems, rapid technological advancement, etc affect both male and female access to and use of ICTs. This may be because while men seek to fully develop and integrate themselves with the technological innovations thereby optimising access to and use of ICTs, women, on the other hand, are not at par with their male counterparts in terms of the necessary skills to utilise ICTs effectively. In as much as equal opportunities abound for both males and females, this difference in factors motivating males toward access to and use of ICTs can be accounted for by female' low self-esteem in ICT use (Sanda & Kurfi 2013; Okunna, 2000) and the impact of culture in the girl child education. Unlike the observed difference in the enablers of access to and use of ICTs among males and females, the current study found no gender difference in constraints to access and use of ICTs. People within same settlement patterns (i.e. rural-rural, urban-urban), employment/job status and requirement, religion/culture, educational level as well as same social strata are affected by same factors such as low level of education, poor skill set, inadequate technological infrastructure, social and economic problems, rapid technological advancement, etc. in their access to and use of ICTs (Sanda & Kurfi, 2013; Morahan-Martin, 2000). Economically buoyant individuals have a higher tendency to cope with technology change, job requirements that necessitate the acquisition or upgrade of their skill set to cope with evolving ICTs and so much more, whether males or females. On the other hand, people with low-income earnings, low level of education as well as poor skill sets are always faced with same constraints such as inability to afford needed ICTs, lack of skills to operate these technologies even when provided, etc.

Future Research Directions

There are numerous studies on equal rights and full participation of women in all spheres of life. However, the gender divide is still dominant among disadvantaged groups, especially those with low-income and low level of education. Given that women constitute the majority of national populations, most of whom live in rural and peri-urban areas, there is a need to assess how the gender difference plays out among disadvantaged groups. Findings from studies of this nature will help in developing appropriate interventions for women in these groups, especially in the area of ICT use in agriculture, health, education, and others; as well as interventions in science, technology, engineering, and mathematics (STEM).

CONCLUSION

This study has shown that there is a disappearing difference in access to and use of ICTs between males and females. Although females are embracing the use of ICTs at a rapid rate, uptake among males still tends to be higher, albeit not statistically significant. Brussevich et al. (2018) note that on average, women perform more routine tasks than men across all sectors and occupations and that such tasks are most prone to automation. As such, fostering gender equality and empowerment in the changing landscape of work remains an imperative such as endowing women with the requisite skills through investment in women in STEM fields. Also, there is a need to bridge the digital divide through access to and use of new technologies to close gender gaps to improve women's labour market prospects in the new world of work. The enablers of access to and use of ICTs tend to be different among males and females. To deepen ICTs adoption and use among females, government's ICT policies and programmes must address the needs of women. While prior literature on gender difference in access to and use of ICTs agree that there is a disequilibrium despite the growing awareness of the benefits derivable from these technologies, this study has shown otherwise. An assessment of the level of inequality in gender's access and use of ICTs is germane to ascertain the extent of its impact (if any) and the causes of a gender-driven digital divide in order to identify measures of attaining equilibrium between males and females. This study has provided insight into the measures that need to be taken in addressing issues of gender inequality in access to and use of ICTs. It is, therefore, recommended that further research effort should be made using a larger sample. Also, a study of this nature that focuses on gender access to and use of ICTs in the public sector will help in the modernising of government through the effective deployment for serving better public sector service delivery.

REFERENCES

- African Gender Institute. (2003). Gender and the Media, African Gender Institute. News Letter: University of Cape Town, 2.
- Agbonlahor, R. O. (2005). Gender Differentials in the Adoption and Use of Information and Communication Technologies by Lecturers in Nigerian Universities. In J. Archibald, J. Emms, F. Grundy, J. Payne, & E. Turner (Eds.), *The Gender Politics of ICT*. Queensway: Middlesex University Press.

- Asiabaka, I. P. (2010). Access and Use of Information and Communication Technology (ICT) for Administrative Purposes by Principals of Government Secondary Schools in Nigeria. *Researcher*, 42-50. <http://www.sciencepub.net/researcher>
- Bhandari, A. (2019). Gender inequality in mobile technology access: The role of economic and social development. *Information Communication and Society*, 22(5), 678–694. doi:10.1080/1369118X.2018.1563206
- Brussevich, M., Dabla-Norris, M. E., Kamunge, C., Karnane, P., Khalid, S., & Kochhar, M. K. (2018). *Gender, Technology, and the Future of Work*. International Monetary Fund. doi:10.5089/9781484379769.006
- CERI. OECD. (2010). Are the New Millennium Learners Making the Grade? Technology Use and Educational Performance in PISA. Paris: CERI.
- David-West, O., & Nwagwu, I. (2018). SDGs and Digital Financial Services (DFS) Entrepreneurship: Challenges and Opportunities in Africa's Largest Economy. In *Entrepreneurship and the Sustainable Development Goals* (pp. 103-117). Emerald Publishing Limited.
- Dominick, J. (2002). *The Dynamics of Mass communication: Media in the Digital Age*. McGraw Hill.
- Dy, A. M., Marlow, S., & Martin, L. (2017). A Web of opportunity or the same old story? Women digital entrepreneurs and intersectionality theory. *Human Relations*, 70(3), 286–311. doi:10.1177/0018726716650730
- Gillwald, A., Milek, A., & Stork, C. (2010). *Towards Evidence-Based ICT Policy and Regulation: Gender Assessment of ICT Access and Usage in Africa, Volume One, 2010 Policy Paper 5*. www.researchICTAfrica.net
- Hafkin, N., & Taggart, N. (2001). *Gender, Information Technology, and developing countries: an analytical study, Learn Link Project, Academy for Educational Development*. http://learnlink.aed.org/Publications/Gender_Book/pdf/Gender_Book_Photos.pdf
- Howell, C., & Lundall, P. (2000). Computers. In *Schools: National Survey of Information Communication Technology in South Africa*. Education Unit: University of Western Cape. <https://www.education.gov.za/LinkClick.aspx?fileticket=nX6%2F4yb8GmE%3D&tabid=106&mid=406>
- Huyer, S., & Hafkin, N. (2007). *Engendering the Knowledge Society: Measuring Women's Participation*. Montreal: Orbicom, NRC Press.
- Kose, T. (2019). Gender and Use of E-Government Services in Turkey: E-Government in Turkey. In *Gender Gaps and the Social Inclusion Movement in ICT*, (pp. 130-146). IGI Global.
- Kuschel, K., Labra, J. P., & Díaz, G. (2018). Women-Led Startups and Their Contribution to Job Creation. In A. Presse & O. Terzidis (Eds.), *Technology Entrepreneurship. FGF Studies in Small Business and Entrepreneurship* (pp. 139–156). Springer. doi:10.1007/978-3-319-73509-2_7
- Kuschel, K., & Lepeley, M. T. (2016). Women start-ups in technology: Literature review and research agenda to improve participation. *International Journal of Entrepreneurship and Small Business*, 27(2/3), 333–346. doi:10.1504/IJESB.2016.073995
- Madanda, A., & Kabonesa, C. (2007). Challenges to Women's Empowerment through ICTs: The Case of Makerere University. *Agenda (Durban, South Africa)*, 71, 81–88.

- Manago, A. M., & Pacheco, P. (2019). Globalization and the Transition to Adulthood in a Maya Community in Mexico: Communication Technologies, Social Networks, and Views on Gender. *New Directions for Child and Adolescent Development*, 2019(164), 11–25. doi:10.1002/cad.20273 PMID:30865363
- Morahan-Martin, J. (2000). Women and the Internet: Promise and Perils. *Cyberpsychology & Behavior*, 3(5), 683–691. doi:10.1089/10949310050191683
- Nsibirano, R. (2006). *Access and use of ICT facilities on Makerere University campus: a mini survey*. Academic Press.
- Nsibirano, R. (2009). “Him and Her” -Gender differentials in ICT uptake: A critical literature review and research agenda. *International Journal of Education and Development Using Information and Communication Technology*, 5(5), 33–42.
- Okunna, S. (2000). *Gender and Communication in Nigeria: Is this the Twenty First Century?* A Paper presented at the 2000 Summit, Barcelona, Youth, Children and women.
- Olatokun, W. M. (2009). Analysing Socio-Demographic Differences in Access and Use of ICTs in Nigeria Using the Capability Approach. *Issues in Informing Science and Information Technology*, 6, 479–496. <http://iisit.org/Vol6/IISITv6p479-496Olatokun628.pdf>
- Olsson, T., Samuelsson, U., & Viscovi, D. (2019). At risk of exclusion? Degrees of ICT access and literacy among senior citizens. *Information Communication and Society*, 22(1), 55–72. doi:10.1080/1369118X.2017.1355007
- Orser, B. J., & Riding, A. (2018). The influence of gender on the adoption of technology among SMEs. *International Journal of Entrepreneurship and Small Business*, 33(4), 514–531. doi:10.1504/IJESB.2018.090341
- Osabuohien, E. S., & Karakara, A. A. (2018). ICT usage, mobile money and financial access of women in Ghana. *Africagrowth Agenda*, 15, 14–18.
- Partey, S. T., Dakorah, A. D., Zougmore, R. B., Ouédraogo, M., Nyasimi, M., Nikoi, G. K., & Huyer, S. (2018). Gender and climate risk management: Evidence of climate information use in Ghana. *Climatic Change*, 1–15.
- Sanda, H. U., & Kurfi, M. H. (2013). Gender and Information Communication Technologies (ICTS) in Nigeria: Challenges and Prospects. *Global Journal of Human Social Sciences*, 13(6). https://globaljournals.org/ev/GJHSS/Earlyview_GJHSS_%28C%29_Vol_13_Issue_6.pdf
- Scanlan, S. (2004). Women, Food Security, and Development in Less-industrialized Societies: Contributions and Challenges for the New Century. *World Development*, 32(11), 1807–1829. doi:10.1016/j.worlddev.2004.05.009
- Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students’ ICT literacy. *Educational Research Review*, 27, 205–217. doi:10.1016/j.edurev.2019.03.007
- Sorenson, K. H. (2002). *Love, Duty and the S-Curve: An Overview of Some Current Literature on Gender and ICT*. http://www.rcss.edu.ac.uk.sigis/public/SIGIS_D02_Part1.pdf

Stewart, J. (2004). Gender as a Factor in the Uptake and Use of ICTs on Family Farms in Southern Queensland, Australia. *Gender, Technology and Development*, 8(1), 97–117. doi:10.1080/09718524.2004.11910106

Tiamiyu, M. A. (2003). Information and Communication Technology for Social Development: Issues, Options and Strategies. In E. O. Soola (Ed.), *Communication for Development Purposes*. Kraft Books Ltd.

Tomte, C. (2011). Challenging Our Views on ICT, Gender and Education. *Nordic Journal of Digital Literacy*, 6, 309–324. https://www.idunn.no/ts/dk/2011/special_issue/art06

Umukoro, I. O., & Tiamiyu, M. A. (2017). Determinants of e-library services' use among university students: A study of John Harris Library, University of Benin, Nigeria. *Journal of Librarianship and Information Science*, 49(4), 438–453. doi:10.1177/0961000616653176

Vernon, R. (2001). *Knowing where you are going: Information System for Agricultural Research Management*. International Service for National Agricultural Research.

ADDITIONAL READING

Adeyinka, F., & Aluko, Y. (2018). Information and Communication Technologies (ICTs), Empowerment and Gender Equity. *Gender and Development in Nigeria: One Hundred Years of Nationhood*, 161.

Agbatogun, A. (2010). Self-concept, computer anxiety, gender and attitude towards interactive computer technologies: A predictive study among Nigerian teachers. *International Journal of Education and Development using ICT*, 6(2), 55–68.

Bakari, D., Gisilanbe, S. A., & Samuel, P. (2018). Assessment of Accessibility of Information and Communication Technology by Gender among Famers in Taraba State, Nigeria. *Archives of Current Research International*, 1–6.

George, T. O., Onwumah, A. C., Fagbohun, M. O., Adebayo, M. E., & Olonade, O. Y. (2019). Gender Differences in ICT Studies: A Study of Selected Public Secondary Schools in Ogun State, Nigeria. In *Gender Gaps and the Social Inclusion Movement in ICT* (pp. 147–169). IGI Global.

Motilewa, D. B., Onakoya, O. A., & Oke, A. (2015). ICT and gender specific challenges faced by female entrepreneurs in Nigeria. *International Journal of Business and Social Science*, 6(3), 97–105.

Nakafeero, A., Okello, D., Comfort, K., Mandlate, J., Kyomuhendo, G. B., Yitamben, G., Tafnout, M., Sifuna, O., Hassanin, L., Macueve, G., & Bakesha, S. (2013). *African women and ICTs: Investigating technology, gender and empowerment*. Zed Books Ltd.

Olatokun, W. M. (2017). Availability, accessibility and use of ICTs by Nigerian women academics. *Malaysian Journal of Library and Information Science*, 12(2), 13–33.

Singh, S. (2017). Bridging the gender digital divide in developing countries. *Journal of Children and Media*, 11(2), 245–247. doi:10.1080/17482798.2017.1305604

Wilson, J. (2016). *Overcoming Gender Inequalities through Technology Integration*. IGI Global. doi:10.4018/978-1-4666-9773-7

KEY TERMS AND DEFINITIONS

Access to ICTs: This is an individual's unrestricted ability, right, or permission to locate and use of an information and communication technology device such as computers, mobile phones, the internet and the likes for the receipt, processing, storage, retrieval, consumption, and dissemination of information.

Digital Divide: This is the gap between technological "have" and "have not" individuals, geographic locations or demographic groups.

Gender: This is a social construction that refers to the roles developed by male and female in response to their environment, especially concerning socio-cultural differences rather than biological ones.

Gender Difference: This is the difference in males and females' mental functions and behaviours that affect their access to and use of information communication technologies as a result of a complex interplay of biological, developmental, and cultural factors.

Information: This is defined as processed data and facts to which meanings have been assigned for the purpose informed decision making that results in a change in behaviour or action as a result of its use or consumption.

Information and Communication Technology: This is a term used for describing all technologies that facilitate the processing, access to, storage, manipulation, use and exchange of information between a sender and receiver of information across various media.

Information and Communications Technology for Development (ICT4D): This refers to the use of technology for helping poor and marginalised individuals and communities in developing countries to bridge the digital divide and aiding economic development by ensuring equitable access to up-to-date communications technologies.

Use of ICTs: This is the utilisation of information and communication technologies for accessing and consuming information and information services consequent of affordability and skills required to manipulate ICTs for strategic advantage efficiently.

How Mobile Technologies Are Leading to Economic Development in Sub-Saharan Africa

Nigel McKelvey

Letterkenny Institute of Technology, Ireland

Adam Crossan

Letterkenny Institute of Technology, Ireland

Kevin Curran

Ulster University, UK

INTRODUCTION

Mobile technology today is increasingly being used to help improve under developed and developing areas such as sub-Saharan Africa. With the statistics showing the number of adults in Africa owning mobile devices steadily increasing, mobile technology has been a popular area of interest to use to help improve areas such as healthcare and education throughout African cities and rural areas. Common that ways mobile technology is being used to help the residents of sub-Saharan Africa are in the sectors of health care and education. Mobile technologies being used in these areas whilst simple are incredibly effective and successful in helping to better the quality of health in education in sub-Saharan Africa. Many of the projects and systems developed using mobile technology focus mainly on urban areas. While reports state the huge increase in the number of those using mobile devices in Africa, along with the large estimated increase in the coming years, the difference between rural African countries and countries which contain urban cities is substantial.

Other uses of mobile technology in African cities includes topping up smart cards with mobile devices for the distribution of clean water as well as the delivery of vouchers which can be exchanged for healthcare equipment. This chapter will show how mobile devices are being used to track and monitor the outbreak of malaria in sub-Saharan Africa and is applied to a system to help ensure health care facilities in areas with affected residents have the capacity to deal with those who have contracted the disease. We also outline the ethical effects of using the mobile devices and how they are affecting the residents of sub-Saharan Africa.

BACKGROUND

Mobile Technology is a sweeping term covering many different areas of technology. Many of the uses for this technology in developed countries include helping with general day to day operations for the user such as online banking, shopping, or social media applications. Mobile technology is a sector which is growing rapidly, and also merging with other sectors to help make use of the technologies full potential. One of the areas where mobile technology has been proven to be successful, is when using

DOI: 10.4018/978-1-7998-3479-3.ch118

simple technology and mobile devices to help developing countries such as Africa (Aker & Mbiti, 2010). Many different organizations and companies are using mobile technology to accomplish huge feats for rural African villages (Abuka, 2016). A leading success story has been the M-Pesa service launched in 2007 by Vodafone for Safaricom and Vodacom. This has generated mobile money revenues of more than \$300 million in 2004. This makes it the most successful mobile transfer service at this time. It has now expanded to Afghanistan, South Africa, India and Eastern Europe. It seems to have found its niche market. In Kenya alone, mobile money transactions totalled \$22+ billion in 2013. Other African countries such as Sudan, Somalia, Tanzania, and South Africa are also huge revenue generators for operators as well (BBC Africa, 2015).

In today's world when discussing developing countries, the most commonly mentioned country is Africa. Africa is known worldwide for its widespread poverty, levels of education, unsanitary living conditions, gang violence and high mortality, and birth rates. Many companies and organizations are now turning to mobile technology to help combat these issues facing those living in Africa. With the increase in advancements in technology, many organizations and companies are researching and looking into how some technologies could be adapted to help reach areas which are usually not easily accessible. A simple yet effective method at developing communication with these areas is through mobile devices. In recent years the ownership of mobile phones in Africa have increased rapidly, with the percentage of adults who own a phone rising from 64% in 2002 to now match the US with 89% in 2014 (Global, 2015). An infographic depicting this information can be seen in the appendices chapter as image 1.0. As shown in reports conducted by Pew Global (2015), the most popular activity for mobile phone users in Africa is sending text messages. The report shows that 80% adult cell phone owners use their phone mostly to send text messages as seen in infographic image 2.0 in the appendices chapter below. SMS is arguably the most popular activity in this area due to the fact that many in this area only have ownership of a cell phone rather than smartphone, with a study conducted on Nigerian residents showing that although 89% of the population there owned a mobile phone only 27% owned a smartphone (Pew Global, 2015).

MOBILE TECHNOLOGY IN AFRICA

Africa is known for its lack of proper infrastructure, and industrialization. Technology is not something that is commonly associated with Africa. The main contributing factor for this being that mostly all technology needs to be powered in some way. A study by The World Bank (2013), showing that 76% of sub-Saharan Africa is without access to electricity, the possibility of being able to use technology within these areas decreases drastically. Another area of technology which is obstructing the growth of use of technology in Africa is connectivity. A study by iPass, (2014) shows that if growth in connectivity hotspot numbers continue to grow at the same rate as previous years then by 2018 there will be a WiFi beacon for every 20 people on earth, however only one beacon for every 400 Africans.

Education

The use of mobile technology has been hugely influential in the improvement of education throughout Africa (Asongu et al., 2016). The need for helping improve education throughout sub-Saharan Africa has been a main priority for many organizations, one of which was Oxfam. Oxfam are a charity organization which help in delivering medical aid, supplies and help to a number of different countries including Africa (Oxfam, 2016). One particular project which Oxfam began was to help improve sub-Saharan

resident's knowledge on different diseases which were quite common in the areas however easily preventable. Oxfam developed an SMS delivery system, which would send out four texts daily with educational information on the different diseases and how to prevent against them. This was incredibly simple as the only technology needed for the residents was a simple mobile phone which could send and receive SMS.

Health

Mobile technology is also being used to help better the health services in areas of sub-Saharan Africa (Asongu et al., 2018a). Organizations and companies such as UNICEF are using simple mobile technology to help teach about how to prevent from contracting diseases, help improve distribution of medical supplies and equipment to people who would otherwise not have access to the materials. They are also helping to monitor and ensure medical facilities contain the stock levels needed to combat against diseases. Two of the main diseases technology had been used to battle the spread of was polio and malaria.

Polio

Alongside the project discussed earlier to help better educate areas of Africa about the polio and how it can be contracted, a project was also launched that would help provide the equipment and sanitary supplies needed to help prevent the spread of the disease to rural and hard to reach areas. The project would also work through SMS where the users would receive a SMS with a mVoucher. As stated by Textello (2016), the mVoucher or mobile voucher is a unique code, each code is created using the customers contact number, resulting in each customer and code being linked, this then reduces the problem of multiple claims, and users mVouchers being stolen. The user would then present their code to a registered trader, where they would then in return receive the specified package containing the items and supplies needed (Oxfam, 2015).

Malaria

To help combat against the threat of malaria in different areas of Africa another project also using SMS was created (Asongu et al., 2018b). This project was however to help survey the outbreak of the disease, and help get real time accurate updates on the impact of the disease. A system to combat malaria such as this was needed as malaria is still a large threat, particularly to children as indicated by the World Health Organization (2015), detailing that a child dies every two minutes in sub-Saharan due to malaria. Rather than simply just supplying information on the disease and vouchers for equipment to prevent from contracting the disease like done with the polio SMS system, the new system would “improve access to essential malaria medicine in areas of developing countries” (Novartis, 2014). The new SMS system created was called “SMS for life”. SMS for life uses common technology to help achieve its goals of eliminating medical facilities from depleting stock levels completely, increase access to essential medicines needed for those infected with the disease and to help reduce the amount of deaths caused by malaria.

SMS for life is a web based system which makes use of a SMS management tool, this then links with a web based application created for reporting (Yeboah-Asiamah et al., 2018). The system uses the SMS management tool to store the number of a health worker from each of the medical facilities registered in the area, SMS for life then uses a short code number that allows the health facility employee to receive an SMS with a data request, and send an additional SMS with the requested data counts completely free. The data sent from the health worker is then manipulated by the online reporting system, which

makes it available for viewing online by a higher body, usually the District Health Management Team. The higher body view this information on the SMS for life website by password, this allows a secure method of monitoring the malaria surveillance in each medical facility in the area (Githinji, et al., 2014).

Water ATM's

In some areas in Africa, local governments have teamed with Grundfos, a Danish water engineering company to install “water ATM’s”. The water ATM’s work by have having a water dispenser installed with users having an account with the dispenser. Each user then receives a smart card, these smart cards and accounts are topped up through mobile phones. The water dispensers can then be activated by the smart card. The user simply scans or swipes the smart card to be dispensed clean, safe water (Wesangula, 2016).

This project has brought water to many different areas of rural Africa and has now began to be installed in more urban areas such as Nairobi (BBCAfrica, 2015). As stated by Mourdoukoutas (2015), there are still some issues with the systems however Rasoul Mikkelsen, Grundfos director for global partnerships believes that “he project will have a huge positive impact on health. It is an affordable system with sustainable, good quality water,”.

As well as the projects discussed above there have also been systems developed using SMS to help monitor Zambia for measles outbreaks, allow HIV sufferers health workers to send information on patient status as well as inform their patients details of their anti-retroviral therapy (Aker & Mbiti, 2010).

SOLUTIONS AND RECOMMENDATIONS

It is clear to see that the results of the Polio projects mentioned above have been extremely successful. An evaluation report by Oxfam (2015) stated that the project had a reach of 104,358 residents receiving the text which in turn was estimated with a reach of 1 million people, this was a completion rate of 86%. These people are now much more educated on how to better protect themselves against Polio, giving not just those who received the informative SMS messages but also those who were in communication with them, a better standard of living and a greater opportunity at life. Also with the pairing of the mVouchers for sanitary supplies and equipment to prevent contraction, with the educational messages it gives an incentive, as it gave the participants of the scheme something physical alongside the messages, something that they can then use the information received to put into practice (Oxfam, 2015).

However, there were some issues surrounding the project. The mVoucher system feedback report shows that the amount of participants that received the voucher compared to the amount of vouchers collected varies greatly with 71.4% of those who received a voucher declining to redeem it (Oxfam, 2015). Oxfam distributed a household survey to each of those who participated in the project, the survey asked the participants if they had collected their supplies with the voucher and if not then why? The main reason given for not claiming the supplies was “Distance to distribution Centre”, this highlights how these projects seem to be specifically targeted at more urban areas rather than the outer lying rural areas of sub-Saharan Africa which is usually in greater need of the aid. This can also be seen in the number of distributors in each district, when the project began it was outlined that each district was to have 3 traders and 15 distributors, however this did not happen in every district, because of the lack in distributors the resulting distribution was below expected capacity, with fewer distribution points (Oxfam, 2015).

Issues can also be seen with the water ATM's which have been implemented in certain cities throughout Africa. Although the ATM's have achieved their goal of delivering safe clean water to different parts of the city, they too can be seen to be focused on the urban areas. As discussed in an article by Mourdoukoutas (2015), it states that Mr. Gichuki, the managing director of the Nairobi Water and Sewerage Company, said that "the Nairobi Water and Sewerage Company has so far identified 611 more points to install the dispensers and aims to eventually install one every 100 meters throughout the city". Meaning the residents from more rural secluded areas are left having to travel further for clean water when those who live in the cities have multiple choices for where to receive their water from, as well as having to wait longer before the installation of a water ATM in a nearer location. This again shows the imbalance between the urban and rural parts of sub-Saharan Africa.

FUTURE RESEARCH DIRECTIONS

The effects of the increased use of mobile technology can be seen in many areas, and with numbers showing that Africa has had the fastest growing region in the years 2009 to 2014, and as stated in a report by GSMA intelligence (2014), is predicted to continue to grow at a rate of 7% annually until 2020. These figures show the level of growth expected for Africa however when looking at figures relating to each individual country a drastic imbalance can be seen. As shown in the report by GSMA intelligence (2014), the breakdown of the number of subscribers to mobile phone providers show the harsh difference between the more urban countries and the more rural countries. For example, the percentage of subscribers in Mali is 70%, compared to the considerably smaller percentage of 22% in Ethiopia. This large margin shows the growing issue of the difference between these two areas, an issue which needs to be examined. As stated in a report by Aker & Mbiti (2010), more examination of the possible impacts that could occur is needed, the report suggested that "careful evaluations of mobile phone development projects are required for better understanding of the economic and social outcomes".

CONCLUSION

From researching and examining how mobile technology has helped improve sub-Saharan Africa, in the different areas of health and education, it is clear that the technology and the projects it was used in has greatly benefited the African residents. The educational system created to aid those at risk of polio helped residents learn cleaning procedures and how to use the sanitary items provided with the mVouchers also distributed through mobile devices also (Okello et al., 2018). The use of mobile devices has also helped service those suffering from malaria, as it ensures that those needing treatment's local health facility will always be within capacity, thanks to the "SMS for Life" system. Finally, the statistics outlining the difference from the rural and urban areas of sub-Saharan Africa, show how the effect differs from region to region. To conclude the chapter, it is evident from the information presented that mobile technology has greatly helped improve different areas of life such as education and health (Porter et al., 2016). However conversely the effects of the increased use of mobile devices in Africa has contributed to a poverty imbalance and perpetuating health inequalities between the urban and rural areas of sub-Saharan Africa, and its possible impacts is an area that needs to be greatly researched before any implementation of larger projects can begin.

REFERENCES

- Aker, J. C., & Mbiti, I. M. (2010). Mobile Phones and Economic. *The Journal of Economic Perspectives*, 22–23.
- BBC Africa. (2015). *Kenya slum Mathare gets cheap water through ATMs*. Available at: <http://www.bbc.com/news/world-africa-33223922>
- Githinji, S. (2014). Using mobile phone text messaging for malaria surveillance in rural Kenya. *Malaria Journal*, 13(107), 1–8. PMID:24642130
- Global, P. (2015). *Cell Phone Ownership Surges in Africa*. Available at: <http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/africa-phones-7/>
- Global, P. (2015). *Cell Phones in Africa: Communication Lifeline*. Available at: <http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/>
- Global, P. (2015). *Few Own Smartphones in Africa, But Cell Phones Common*. Available at: <http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/africa-phones-5/>
- GSMA Intelligence. (2014). *The Mobile Economy - Sub-Saharan Africa*. London: GSMA.
- iPass. (2014). Available at: <http://www.ipass.com/wifi-growth-map/index.html>
- Mourdoukoutas, E. (2015). *ATM-operated water dispensers bring safe water to poor communities*. Available at: <http://www.un.org/africarenewal/magazine/december-2015/africa-wired>
- Novartis. (2014). *Innovation home*. Available at: <http://malaria.novartis.com/innovation/sms-for-life/index.shtml>
- Oxfam. (2015). *Oxfam Evaluation Report*. Oxford: Oxfam GB.
- Oxfam. (2016). *Our purpose and beliefs*. Available at: <https://www.oxfam.org/en/our-purpose-and-beliefs>
- Textello. (2016). *mVoucher*. Available at: <http://www.textello.com/features/mvoucher/>
- The World Bank. (2013). *The World Bank and Energy in Africa*. Available at: <http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/AFRICAEXT/0,contentMDK:21935594~pagePK:146736~piPK:146830~theSitePK:258644,00.html>
- WesangulaD. (2016). *Mail & Guardian*. Available at: <http://mg.co.za/article/2016-02-18-nairobi-slum-gets-water-from-atms>
- World Health Organization. (2015). *African populations are extremely affected by malaria*. Available at: <http://www.afro.who.int/en/clusters-a-programmes/dpc/malaria/features/2287-10-facts-on-malaria-in-africa.html>

ADDITIONAL READING

- Abuka, C. (2016). *State of financial inclusion in Uganda: Bank of Uganda Policy Report, 2016*. Bank of Uganda.

Hellström, J. (2010). *The Innovative Use of Mobile Applications in East Africa*. Swedish International Development Cooperation Agency.

Horvath, T., Azman, H., Kennedy, G. E., & Rutherford, G. W. (2012). Mobile phone text messaging for promoting adherence to antiretroviral therapy in patients with HIV infection. [Google Scholar]. *Cochrane Database of Systematic Reviews*, 3. PMID:22419345

Jennings, L., & Gagliardi, L. (2013). Influence of mHealth interventions on gender relations in developing countries: A systematic literature review. *International Journal for Equity in Health*, 12. PMID:24131553

Kunutsor, S., Walley, J., Katabira, E., Muchuro, S., Balidawa, H., Namagala, E., & Ikoona, E. (2010). Using mobile phones to improve clinic attendance amongst an antiretroviral treatment cohort in rural Uganda: A cross-sectional and prospective study. *AIDS and Behavior*, 14(6), 1347–1352. doi:10.1007/10461-010-9780-2 PMID:20700644

Leon, N., & Schneider, H. (2012). *MHealth4CBS in South Africa: A Review of the Role of Mobile Phone Technology for the Monitoring and Evaluation of Community Based Health Services*. Medical Research Council, University of Western Cape.

Okello, G., Bongomin, C., Ntayi, J., Munene, J., & Malinga, C. (2018). Mobile Money and Financial Inclusion in Sub-Saharan Africa: The Moderating Role of Social Networks. *Journal of African Business*, 19(3), 361–384. doi:10.1080/15228916.2017.1416214

Porter, G., Hampshire, K., Milner, J., Munthali, A., Robson, E., De Lannoy, A., Bango, A., Gunguluza, N., Mashiri, M., Tanle, A., & Abane, A. (2016). Mobile Phones and Education in Sub-Saharan Africa: From Youth Practice to Public Policy. *Journal of International Development*, 28(1), 22–39. doi:10.1002/jid.3116

Traxler, J. (2005). The Use Of Mobile Learning To Support Teacher Education. In *Sub-Saharan Africa*. Department for International Development.

Yeboah-Asiamah, E., Narteh, B., & Mahmoud, A. (2018). Preventing Customer Churn in the Mobile Telecommunication Industry: Is Mobile Money Usage the Missing Link? *Journal of African Business*, 19(2), 174–194. doi:10.1080/15228916.2018.1440462

KEY TERMS AND DEFINITIONS

Advance Fee Scam (Nigeria 419): Is an advance fee fraud which involves a payment in advance with the promise of wealth, wills, gifts, employment or prizes. The scammer contacts their intended target via mail, phone, fax, or email. Originally known as the Spanish Prisoner Letter it has been carried out since at least the 16th century via ordinary postal mail. Section 419 of the Nigerian Penal Code prohibits this type of activity and it has often referred to as ‘419 frauds’ or ‘Nigeria 419’.

Developing Countries: Are often referred to as underdeveloped countries and they refer to countries where there is a less developed industrial base and they have a low Human Development Index (HDI). Developing countries differentiate from developed countries in that the people have a lower life expectancy, lower standard of education and the people of developing countries have a lower income and thus less money.

Ethical Standards: Are principles which when followed will promote values such as trust, good behaviour, fairness, and/or kindness. One consistent set of standards that all companies follow does not exist but rather each company has the right to develop a set of standards which are meaningful to that particular organization.

Information and Communication Technology (ICT): Is an umbrella term which includes any communication device or application, encompassing: radio, television, cellular phones, computer and network hardware and software, satellite systems and so on, as well as the various services and applications. The term ICT however has no set definition as the concepts methods and applications involved in ICT are constantly changing almost on a daily basis.

Mobile Point of Care (MPOC): Involves the use of multitasking devices which support clinician use cases, improve productivity, streamline communication and bridge the gap between inpatient and outpatient care. This fast growing area provides an abundance of apps for mobile devices and will change the way healthcare is delivered providing the opportunity for remote monitoring, measuring drug response, tools for drug development and diagnostic assistance for point-of-care and management of chronic or infectious disease.

Mobile Technology: Is a technology which is used for the purpose of cellular communication. The standard mobile device has moved from being a simple two-way pager to being a mobile phone. Mobile technology as the name implies refers to a technology which is portable. Examples of mobile IT devices include laptops, netbook computers, tablets, smart phones, global positioning system (GPS) devices.

Monitoring Applications: Involve the use of a technology to enable the monitoring of patients while outside of conventional clinical settings (e.g., in the home), which may increase the access to care and decrease the cost of delivering healthcare.

Personal Data: Refers to data relating to a living individual who is or can be identified from the data or using the data in conjunction with other information which is in or likely to come into the possession of the data controller.

Short Message Service (SMS): Is a simple text messaging service component found in phones, the Web, or in mobile communication systems. It adopts a series of standardized communications protocols which enables either fixed line or mobile phone devices to exchange short text messages. More commonly referred to using the term “SMS” it is used for both user activity and all types of short text messaging in many parts of the world.

Sub-Saharan Africa: Geographically refers to all the areas of the continent of Africa which lie south of the Sahara Desert. Some of the countries which fall into this category include countries such as Kenya, Ethiopia, Ghana, Nigeria, and South Africa.

Section 14

Library Science and Research Methods

Informational Competencies

José Poças Rascão

 <https://orcid.org/0000-0003-2448-2713>

Polytechnic Institute of Setúbal, Portugal

INTRODUCTION

The information at the end of the twentieth century became an important factor of production, although it was always present in the history of mankind, but it did not always have the same importance as it has in the post-capitalist society. In the Middle Ages people lived from what the earth produced, cultivated it and lived according to the productive cycles of the seasons.

The craftsman lived and had the workshop in the same place. It produced the works according to customers' requests, producing the final products (their art). With the Industrial Revolution, the worker began to live in function of the clock and the completion of tasks, one after another, until the whistle of the factory announced the end of the work day. The worker did not need to know what would be done with the result of his work, that is, what was the final product. The worker lost the relationship between the work performed (production) and its result.

Marx separated the worker from the product of his work, at the factory, having invented the work as something separate from the worker. The notion of product and production began to belong to the owners of the factories. In the factory it became necessary to control the production on a large scale, and for this, the production administration was invented. The scientific administration developed the conditions of productivity in an unprecedented way and prepared the mechanization of the tasks before performed by parts of the human body. At the limit, manufacturing production could be fully mechanized and dispense with all workers, minus supervisors. Someone would have to control the machines.

Focus and Approach Methodology

As for its nature, the research is qualitative, since it does not claim to quantify events or privileges the statistical study. Its focus is to obtain descriptive data, that is, the incidence of topics of interest in two fields, the information Sciences and the Sciences and Management. Consequently, about the extremities, the research is exploratory in nature and descriptive, as the technique used, is categorised, consensually, as a study of direct documentation, which provides for the consultation of sources Related to the study in different media, printed or electronic.

BACKGROUND: FUNDAMENTAL CONCEPTS

Information and Knowledge

Although the terms information and knowledge are used very often, they are not the same thing. Information is not the same as data, although the two words are often confused, so it is understood that the

DOI: 10.4018/978-1-7998-3479-3.ch119

subtle distinction between these concepts is essential. Data do not convey meaning or meaning of facts, images or sounds, since they lack relational elements indispensable to the establishment of a complete meaning, lacking an internal relational structure for a cognitive purpose.

This structure is one of the attributes of information. Data becomes information when its creator adds meaning (Davenport and Prusak, 1998). William G. Zikmund (2000, p.19) defines knowledge as “the mixture of information, experience and understanding that provides a framework that can be applied to the assessment of new information or new situations”. Information “feeds” knowledge. Knowledge can thus be defined as a person’s ability to relate complex information structures to a new context.

New contexts imply change, action and dynamism. Knowledge cannot be shared, although the technique and components of information can be shared. When a person internalizes information to the point of being able to use it, we call it knowledge (Zikmund, 2000). This is a fluid mix of structured experiences, values, contextual information and expert insight that provide a framework for evaluating and incorporating new experiences and information. In organizations it is found not only in documents and reports, but also in organizational routines, processes, practices and standards.

Knowledge has its origin and is applied to the minds of connoisseurs (Davenport and Prusak, 1998, William Zikmund, 2000). Knowledge is information as valid and accepted, integrating data, acts, information and sometimes hypotheses. Knowing requires someone to filter, combine and interpret information. Information can be considered as a “substance” that can be acquired, stored and owned by a person or a group and transmitted from person to person or from group to group.

Information has a certain stability and may be better viewed as existing at the societal level (Davenport and Prusak 1998). Although we may store it by employing various physical supports, the information itself is not physical, but abstract and so purely mental. Knowledge is stored in people’s memory, but information is out there in the world. Whatever it is, there is somewhere between the physical world around people and the mental world of human thought.

Knowledge = Internalized information + Ability to Use It In New Situations

Knowledge lies fundamentally and intrinsically within people. These are more complex and unpredictable at the individual level than an entire society, so it is not surprising that knowledge is much harder to obtain than information. Knowledge exists mainly within people; it is an integral part of human complexity and unpredictability.

Knowledge has a fundamental duality: it is something storable (at least sometimes we intend to do so) and something that flows (something that communicates from person to person). It is possibly the duality of knowledge (thing that flows and the storage process) that makes its treatment and management difficult. According to Dahlberg (2006) knowledge is organized into knowledge units (concepts) according to their characteristics (objects / subjects / subjects). Knowledge organization is related to a process of conceptual analysis of a knowledge domain and from there it is structured / archived generating a representation of knowledge about such domain that will be used for the organization of information about that knowledge domain.

Resource

A resource is anything that can be thought of as strength or weakness for an organization, so that it gets a competitive advantage if it acquires or develops superior resources or a combination of superior resources.

Teece, Pisano and Shuen (1997) define how the specific assets of a company's resources are difficult, if not impossible to imitate. In turn, Prahalad and Hamel (1990) define the resource as "nuclear Powers".

A feature incorporates several strategic management skills to explain how some organizations have superior performance to their competitors. This view means that organizations are a set of resources and capabilities in which the first management task is to maximize the value of existing resources and capabilities, while developing the basis of the resources of the Future, (Grant 1991, p. 110).

The distinction between resources and services consists of the following: the resources are a set of services and the potential that can be defined independently of their use, while the services cannot be defined (Penrose, 1959, p. 25). The capacities of an organization at a given time are those of tangible or intangible assets that are connected to the organization as capital, equipment, skills and knowledge of workers, brands and technological Systems (Wernerfelt 1984, p. 172).

However, few resources are productive, but require cooperation and coordination together. This capability of the resource pool to perform some tasks or activities, is referred to as a capacity. In order to make productive use of this set of resources, organizations acquire resources that are called market factors, such as the labor market or capital.

Market strategic factors are defined as the markets in which organizations acquire the resources needed to market their strategic products. Organizations acquire resources as market factors, work and develop resources that make use of these resources in order to supplement the products and services that are sold in one or more markets.

Costs

Cost is an expense relative to a good or service used in the production of other goods or services. Cost calculations have been limited to those associated with product production. Some studies have been directed to the costs of the use of information and others still for the costs of certain channels of information acquisition.

Effectiveness

The evaluation of the efficacy of a service or product presupposes the comparison of objectives with the results. Efficacy is a measure that measures when we achieve a goal or to what extent we have achieved a satisfactory level of performance. The measure of efficacy is related to the maximization of results at satisfactory costs, not necessarily the minimum. The effectiveness of Information Services is seen as a desirable state when the highest level of performance is obtained, consistent with the levels of spending and resources consumed. In the analysis of effectiveness there must be a close relationship between the objectives of the services and the needs of customers (users). A measure of value and cost of each product or service is required for proper evaluation of efficacy. Thus, this analysis is valid to the extent that these items are adequately quantified. The benefits and costs that cannot be directly data, in monetary terms, can be quantified by coefficients with relative weights.

Efficiency

While the study of efficacy compares objectives and outcomes, the study of efficiency focuses on how well the activities are performed. This study has been very common in the evaluation of information services. Efficiency is related to maximising the effectiveness of spending. This implies eliminating

unnecessary expenses for the execution of defined tasks, thus seeking the minimum cost; This will not necessarily maximize performance.

Value

The VALor is an attribute that can be applied to almost anything. The notion of value is sometimes confounded with price and cost. **Price** is the monetary amount to acquire a good or service by its owner; the price defines the monetary amount that can be exchanged for the good or the service. Cost is the monetary amount spent in the production or execution of the good or service. The **value** has three characteristics: it is subjective; varies with time and can be positive or negative. The positive satisfactions of value are called “benefits” and the negative are “loss” or “damage”.

Skills

Information and knowledge became integral factors of production. Other values come into play: to innovate and deal with irregularities, it takes autonomy and responsibility, in addition to a knowledge of processes and products. Information and knowledge have always been part of the productive work, but there is now visibility, and its importance is acknowledged. As knowledge can only be mobilized by the workers, by the human person, the logic is reversed. The work is no longer the factor something external to the worker and is once again something intrinsic to those who accomplish it.

The transformations that occurred in the world of work replayed the human being at the center of production. It then goes on to talk about skills, and no more qualification for a job or a particular job. It is the person, with its most complete characteristics that matters. For a peasant or a craftsman, working meant employing and developing his knowledge, relying on rules transmitted by more experienced peers (Masters) and acquired by the individual through a series of proofs (...) was the direct expression and Immediate practical intelligence. It was the invention of work as a separate object of the worker who served to qualify another type of social relationship, that of the objective work and the capacities necessary for the *accar*. Hence the separation between the objectivity of work and the subjectivity of the worker. To get out of this logic of objectification, it was necessary that the work was again subjectivated. This meant the inclination to the logic of competence, which was called “model of competence”.

Osty (2000, p. 9) considers that the notion of competence shows the emergence of a new mode of management and recognition of the knowledge required by work situations, revealing the breadth of work transformations in the last twenty years of TWENTIETH century.

According to Zarifian (2003, p. 37), competence is a new form of qualification, a new way of qualifying. But it is not about qualification as “a particular and always dominant historical way: that of qualification by the workplace”, but rather in a new way: “The construction of qualification”. This construction of qualification, in the face of new production conditions, would replace the two previous models: the model of the profession and the model of the job (or employment). The model of the profession began to be built in urban craft corporations. The “*metier*” or function, a term used since the middle ages by medieval corporations, presupposed a means of learning and a social pertinence. This model began to be the target of attacks at the end of the eighteenth century because it seems an obstacle to the transformation and rationalization of the manufacturing work methods. It defines skill as “the ability to accomplish a task or a set of tasks, in accordance with certain standards required by the organization”, involving theoretical knowledge and personal skills.

The use of the notion of competence began to appear when companies had the need to recognize people's competences, regardless of the job they were occupying. The emergence of the model Dand Competence represented a new moment, whose practical bases are still being developed. Many alleged competency management systems are still only modernized forms of the workstation model.

For Brandão (1999, p. 22) The meaning of the word competence has been expanding since the end of the Middle Ages. There was, firstly, a legal meaning: "Faculty attributed to someone or to an institution, to appreciate and judge certain questions". It later emerged an extension of the concept: "Social recognition on the ability of someone to pronounce, regarding a specific subject".

Skills Type

Nordhaug (1998, p. 10), classifies skills in: technical, interpersonal and conceptual. The technical competencies are related to the methods and processes, designed to conduct a specific activity and skills to use tools and operate equipment related to an activity. Interpersonal skills are human behaviors and interpersonal processes, empathy and social sensibility, communication skills and capacity for cooperation. The conceptual competencies are analytical skills, creativity, efficiency in solving problems and the ability to recognize potential opportunities or problems.

INFORMATION AS RESOURCE OF ORGANIZATIONS

Information Economics

Economic theory is concerned with answering two fundamental questions: the determination of how scarce resources are distributed in a complex society; The way in which these resources are distributed that maximizes the benefits to society. Many analysis models have been developed to answer these questions. The term economics of information began to appear in the economic literature from 1960 onwards. In the information sciences, it was initially related to the evaluation studies. More recently, in the 1980, the cost and Effectiveness of information services approach was included, followed by discussions on the value of information for the user and the productivity of information in Work.

The literature on the subject has been published in a more pronounced way in the United States and Great Britain. The study Dthe information economy presupposes the definitionDthe concept Dand information. In a popular sense, information is understood as any clarification that can be given to a person about what she asks. Information feeds knowledge in new situations. The information may be shared, but the knowledge is personal and is only shared if the knowledge holder wants to.

Primary data from credible sources are collected, processed, processed, sorted, transferred, evaluated, and so on. The information is viewed as a product or even as a resource, but the information is a unique resource, of specific nature and very characteristic characteristics. Cleveland (1982) characterizes the information as follows:

- **Information is "human"** - there is only information through human observation;
- **The information is multipliable** - the more we use it, the more useful it becomes; The basic limit is the biological age of people and groups;

- **The information is replaceable** - it can replace other features such as money, people, raw material, etc. For example, the accumulation of information in area D Automation replaces several million workers annually.
- **The information is transferable** - the speed and ease with which the information is transferred is a considerable factor for the development of all areas of knowledge;
- **The information is diffusive** - it tends to become public, even if our efforts are in the opposite direction;
- **The information is shareable** - the goods can be exchanged, but, in the exchange of information, the seller continues to possess what he sold.

The study of the information economy has often used classical economic theories. However, during practical studies and experiments, some special features of the information have emerged:

- The information products have value, but their benefits depend on the ability of those who exploit them;
- The information is not consumed with the use; Only time sometimes makes information unimportant;
- Information is not a constant, that is, usually it cannot always be quantified;
- Information is an abstraction, that is, it is produced, disseminated, accumulated and used for different objectives and purposes. This feature causes a lot of confusion, for example, when someone estimates the value of the information according to what the other is willing to pay for it;
- New information is usually produced with public resources (basic research), and these costs are usually not imputed for the purpose of defining the market price;
- The real benefit of the information is difficult to measure because it is limited to its use, which is impracticable.

The study of the information economy becomes problematic in the face of these information characteristics. The information economy is, however, becoming an important object of research in the information sciences.

Economic analysis plays a key role in assessing the cost-benefit ratio. Cost-benefit analysis is a technique that seeks to determine and evaluate the costs and social benefits of investment to help decide whether or not a project should be implemented. The emphasis on costs and social benefits distinguishes it from other investment techniques. The technique tries to measure losses and gains, in terms of economic well-being for society, both in evaluating a project itself and in the different options of taking it forward. The cost-benefit analysis helps decision making when it seeks to answer the following questions:

- How much should the organization invest in information products and services?
- Of the current costs of the organization, how much is it to refer to information services and other activities?
- Are the resources currently budgeted for the information appropriate or should it be altered?
- How are resources allocated between the different information services?
- Are investments in information services successful or should they be changed?
- What use has every project of information service?
- How are prices given to information services?
- How can the production and dissemination of information be supported or taxed?

- Are the costs appropriate to the organization's management and investment plans?
- What measures are necessary to monitor the performance of information services?

Some investigators have criticized a lot and indicated flaws in the generalization of the cost-Benefit analysis of information. They affirm that both the scientists of the information sciences and the economic sciences have doubts about the possibility of measuring the value of the use of information. However, the cost-benefit analysis is potentially appropriate for decisions on the allocation of resources in social welfare projects, such as in libraries and information services where the benefits are not always direct and Tangible.

Companies have undergone a phase of transition from an industrial economy to an information and knowledge economy, that is, an environment where information adds value to products and services. The companies that have been used to succeed, are losing ground for more agile competitors in the use and information management.

In this context, the information is treated as a strategic resource that needs to be administered daily, with the same attention to the human and financial resources of the organizations.

The economy is growing in the production of goods, services and information, and can be generally characterized by the following assumptions:

- The perception that the industries and the economic sectors will be dependent on a workforce that has access and can work on the information;
- The recognition that information needs to be organized, structured, systemized, stored, managed and easily accessible;
- H recognition that the skills of creation, search, analysis, evaluation and interpretation of information are essential for individuals and groups, both inside and outside organizations;
- The perception that information needs (what information?) become increasingly complex and dependent on different and multiple sources, whose correct evaluation and quality is a crucial factor for decision – making processes;
- That the growing development of information and communication technologies challenges both the competences of the Laity, and the Information Professionals, in terms of their understanding, mastery and management;
- Recognition that the information sector is a substantial part of the economy of the country and the World.

Economic Profitability

For economics the need is the sense of a certain lack attached to the desire to be eliminated. The need, therefore, is the desire for a means capable of preventing or ceasing an unpleasant sensation or of provoking, preserving or ameliorating a pleasant sensation.

The media that serves people to meet their needs is designated by economists as goods. A good, all that we recognize as capable of satisfying a need regardless of any other judgment (egg moral judgment). Goods can be classified in several ways. The first thing that interests us is the one that distinguishes the so-called free goods (for example, the air we breathe, water) from economic goods. An economic good is characterized by its scarcity and is one that exists in smaller quantities than required, given the felt need.

The utility of a good refers to the quality and ability of that good to meet the needs that are appropriate for that good. The utility of a good is the ability of that good to satisfy the needs that are subjective

sensations, it follows that the same good may have different utility for different people, depending on the intensity with which they can experience or feel the needs that can be met, for this good.

Closely associated with the concept of utility is the concept of “economic value”, of which we can say that it is the quality of everything that is given importance, because it is considered worthy of esteem. Thus, the value of a given good also depends on each person, since each one may feel distinct appreciation for that good.

Organizations are seeking economic return by selling their products or services. In a competitive environment, an organization can perform better than its competitors because it has better characteristics or because the organization makes better use of its distinctive skills (Penrose, 1959 p. 54). A distinct competency is defined as a distinct qualification, complementary assets, and organizational routines that together enable the company to coordinate a specific set of activities that provide the basis for competitive advantage.

Competitive advantage is referred to as economic profitability (Porter 1980). Subsequently, the resource base views the strategy as an ongoing search for profitability. Profitability is then defined as the excess return on a resource that owns an opportunity cost (it is the return that can be generated when put into alternative use or the price at which it can be sold (Grant, 1991). In other words, profitability can be measured as a normal financial return.

In order to differentiate possible sources of profitability, some types are distinguished. For example, the return may be accomplished by:

- Superior ability to manage and coordinate resources;
- Possession of a valuable resource that is rare;
- Government protection when barriers to entry for new competitors are high;
- Business execution risk in uncertain and complex environments;
- Make better use of the organization’s resources, any of its physical assets and human capital.

The resource base is an attempt to explain why organizations differ in resources and capabilities and how these differences can lead to sustained profitability positions, producing superior financial returns. Core resources, therefore, serve the purpose of explicitly focusing on the role of resources and the capacity of organizations, such as strategy origin and organizational performance. This exploration of the relationship between resources, competition, and profit includes, among other issues, the role of imperfect information in creating profitability differences between competitors (Grant 1991, p. 3).

Information Profitability

Information is a complex concept, ubiquitous since the emergence of life and inherent in any evolutionary process and thus served and continues to serve as a milestone for reflection in both biology, physics, psychology, management, linguistics, as in various other Branches of science.

The power of information resulted from the development of information and communication technologies during the twentieth century, as occurred two centuries ago, with the concept of energy, a key factor of the industrial revolution.

The imperfections of information are some of the competitive imperfections that collide with the economic ideal of the perfect market. The perfect market is characterized by countless customers and suppliers of homogeneous products, whose pricing mechanism determines the most efficient market

organization. Looking at the information, all customers and suppliers are fully and perfectly informed about all relevant aspects of business transactions.

While the ideal market serves as the most important neoclassic economic theory, the appropriations of contemporary economic theories tend to allow this, contrary to the appropriation of unreal theories. Therefore, we assume that in the perfect market, customers and suppliers are perfectly informed, but that is a myth. For this reason, it is proposed that markets are characterised by imperfect competition and that the imperfections of information are an important trend.

The amount of return that an organization can perform depends on the cost of resources, the cost of acquiring new resources, and their economic value, when the products are sold out. In the competitive markets, customers and suppliers will have the same expectations, about the value and cost of resources and products.

Consequently, vendors never sell a resource or their product, if the total value is not reflected in the price, or customers will pay a higher price than the cost. In these circumstances of the perfect market, the strategy can only produce the normal return, as opposed to the return mentioned above, namely profitability.

Information Imperfections

To understand how economic profitability can be identified and appropriated by strategic information, it is through knowledge of the origin of the imperfections of information. The costs of economic transactions (CTE) describe these origins and three types of information imperfections resulting from them.

The theory of transaction costs assigns two behavioral appropriations to man: the limits of rationality and opportunism. We assume that men are intentionally rational, but they are limited by that. The limits of rationality are beyond the inability of the human brain. This involves neuropsychological boundaries of one skill, on the one hand, and on the other, the limits of language. Physical boundaries take the form of value and storage limits of the power of individuals to receive, store and process information, without error.

Human cognitive abilities are generally insufficient to understand particularly complex tasks, to recognise the need for pertinent information and to make all relevant information. It explains the limited rationality of the rational behavior of information. Instead of a thorough search for relevant information, the only decision maker tries to obtain a subjective satisfaction, the level of incomplete information. This type of behavior is known as “satisfactory”: looking for a solution that meets the aspirations of decision maker level and therefore the acceptable.

Human beings are not only rational; they also assume some opportunistic behavior. Williamson (1975) describes opportunism as “the pursuit of self-interest, with cunning” involving “disbelief of threats and opportunities” in order to accomplish the individual advantages. These advantages are also appropriate to “select and distort the discovery of information or discredit the opportunities that lead to the future” (Williamson, 1975, p. 26). The claim is that although not all humans are comportandm opportunistic, it is difficult to predict in advance whether they will behave or not, as such.

These two types of behaviors reflect the relationship between vendors and customers. The combination of opportunistic and rational behavior is the main cause for the three types of information imperfections in the transaction market:

- **Asymmetric information** - an asymmetric distribution of Information, from the parties involved in a transaction, causes a strategic opportunity. The critical impact of information on optimal allocation of risk is not merely its presence or absence, but the inadequacy of its economic agents. Thus, it joins two conditions in which the information asymmetry provides a strategic opportunity in the transactions:

- I. High costs for achieving equal information;
- II. Propensity for opportunistic behavior. For other words, asymmetric information occurs, when one of the parties has information that is unknown to the other and of hard obtaining for the other part. You have an opportunity to explore this advantage of information through strategic alternatives.

Asymmetric information may result in a favorable situation for suppliers or customers, depending on what has this information. The opportunity is caused by hidden information for the transaction. One of the parties in a business transaction is better informed, about a relevant variable than the other. It is the invisibility of this private information that constitutes the essence of the imperfections of information and introduces the risk to the other party.

Considering the possibility of opportunistic behavior, the part that has hidden information, has no incentive to disclose it, if it is harmful to it. As a result, any supplier entering the market with a product or service that is particularly suitable for a specific market segment, with the main benefits for competitors benefiting from the hidden information. The imperfections of information are known as enemies.

- **Ambiguous information** - Information ambiguity in a business transaction can cause a strategic opportunity. Although in the case of complete information, different interpretations about them may occur. As a result of opportunism, these representations can be opportunistic in the sense that it can lead to a single advantage for any supplier or customer.

The strategic opportunity for one of them is discovered, when such opportunistic interpretation is not recognised by the other and results in:

- i. The customer will pay a higher price for a product;
- ii. Supplier availability to offer a low-price product.

As a result of this opportunistic behavior, the part with opportunistic interpretation has no incentive to share this information if it becomes a disadvantage to it.

- **Incomplete information** – the strategic opportunity for incomplete information is closely related to the assumption of rationality. The limits of rationality are certainly interested in expanding the limits of rationality achieved, in conditions of uncertainty and complexity. In the absence of any of these shares' ownership terms can be quite specific to this principle.

In an environment characterized by much complexity and uncertainty, the possibility of certain events occurs quickly, and become numerous. In these circumstances, it is impossible for humans to acquire and analyse comprehensive information relevant to strategic decision making. As explained earlier, this inability leads to satisfactory behavior. The opportunity for incomplete information occurs when

a customer is not fully informed about the range of possibilities of the transaction, about the products offered and vice versa.

Therefore, a client that allows an incomplete comparison of the vendors and their differences, can initiate a transaction with the worst vendor. For the supplier this opportunity can be exploited for the creation of a brand.

Quality of Information

Not all information is good, one is better than another. It is therefore necessary to assess its quality. The criteria for assessing the quality of information are as follows:

- **Relevance** - the information must be *relevant*, that is, it must relate to the facts, be available and be important to the person who requires it. The information will help people make decisions.
- **Opportunity** - information should be *timely*, that is, to be available to the right person at the right time. If, for example, a person doesn't know the train's departure time, they risk losing it.
- **Accuracy** - The information must be *accurate*, this means that if the information is not accurate, it loses interest;
- **Reduction of uncertainty** - just think about taking decision with and without information; good information reduces uncertainty. Good information involves the difference that *makes a difference*;
- **Element of surprise** - information can be used to obtain competitive advantages;
- **Accessibility** - Information is only useful if people have access to this; *Accessibility* is within reach of those who can obtain the information, in time to being used efficiently and in the form that makes them useful. Electronic storage makes information more readily accessible than pencil and paper technology.

RELEVANT ISSUES

Information Needs

To manage the production is necessary information about each worker, his/her tasks, working hours, productivity, equipment, product and its characteristics, etc., but is not yet a factor of production. With high productivity and low costs, the market becomes a salesman instead of a producer, consumers become increasingly demanding increasing competition and begin to influence production: product quality and preferences of Consumers, give rise to the economy of services, on the counterpart of the economy of the product.

In this economy, there are three factors influencing production: the event (or eventual situations), communication and service. Production begins to target events; Which does not conform to the regularity of mechanical production. Whether to cater to different customers with their specificities, be it to solve the normal production problems. Human activity is now due to irregularities, novelties and innovations. In this way, information gains paramount importance in production. You can't handle events and innovations without information and knowledge about an ever-increasing set of operations. Communication also becomes a primordial factor because problems become more complex, requiring more than one area of knowledge to solve them. It is necessary to use the information efficiently so that there are no communication problems, that is, the information becomes one of the factors of production.

Identification of Informational Competencies

For identification the informational competences, can be taken as basis, the various typologies approached previously and, from them, select the specificities. The interpersonal informational competences and the conceptual informational competences may be related to the skills of mediation of the professional who works with the information to make the meeting between information and your user.

In the typologies of Fleury & Fleury (2001) Organizational informational competencies are the set of individual informational competencies embodied in organizational processes, in order to construct specific competences related to the information activities essential to the organization. The key competences of Prahalad & Hamel (1990) could correspond to informational competences developed in an organization with the specific objective of differentiating or improving the competitiveness in the market.

In the typology of Zarifian (2001), General and professional knowledge (references of a given professional universe) are linked to the characteristic training of an information professional and the background informational competences (acquired in Educational situation and formalized in cognitive and behavioral achievements necessary to address the categories of problem situations) are linked to what will be developed or mobilized in specific organizations or activities, to solve the specific informational problems faced in each case.

Four categories of key management competencies can be identified:

- **Relational competencies** - work productively with others;
- **Troubleshooting skills** - identify, formulate and execute creative solutions to problems;
- **Capacity-building skills** - recognizing the need to change and undertake change;
- **Communication skills** - communicate effectively and efficiently.

The skills for problem solving are perception capabilities, capacity planning and organization, and capacity to make decisions. These competencies include the selection of information, the mastery of appropriate methodologies and tools for the treatment of this information, with a view to generating appropriate results, with effectiveness and efficiency. Communication competence involves oral and written communication, focusing on the mastery of communication processes, which involves language, the ability to understand and transmit ideas.

For the purpose of identifying informational competences, it is considered that the competences linked to information are related to the informational process, process that, studied by the information sciences. These competences relate to information and communication technologies, with the cycle and analysis of information. These are competencies that permeate the business processes, the management processes and the various technical processes, as well as different parts of the same organization. Other components of informational competence should be identified in the analysis of situations, by the expression of the initiative, responsibility, transformation and mobilization of actors and sharing.

For any organization the problem is to define the model of competencies, that is, how to organize and structure the information to facilitate communication, given the complexity and turbulence of the world of work, since the communication became a primordial element. The information is what specifies, selects or singularizes the requests, in view of successful professional conduct. What matters in this professional situation is the pertinent information about what someone requests in view of the action, allowing the individual to situate in the Environment and act effectively.

Informational Competence

To deal with informational competencies the following concepts were chosen:

- **Competence** - is "... A knowing act responsible and recognized, which implies mobilizing, integrating, transferring knowledge, resources, skills, which add economic value to the organization and social value of the individual "(Fleury & Fleury, 2001, p. 21). The main attributes of competence are the initiative, Responsibility, Practical Intelligence, acquired knowledge, transformation, diversity, mobilization of actors and sharing (Zarifian, 2001 and 2003).
- **Professional competence** - it is related to people or work teams, integrating technical, cognitive, social and affective aspects related to work (Brandão, 1999, p. 28). It comprises knowledge, skills and attitudes or behaviors that will allow the organization's development, in fulfilling its mission (Dutra, 2001);
- **Organizational competence** - is the *savoir-faire* of the company in a particular domain, which originates and sustains itself by professional competences, allied to the organizational processes and other resources, or the ability to combine, mix and integrate resources into products and / or services (Fleury & Fleury, 2001, p. 23). They include the skills of the organisation and its processes, skills in techniques and forms of work, service skills and social skills;
- **key competence** - is the set of skills and technologies whose brand of authenticity is integration. They represent a perceived value by the customer, a differentiation between competitors, an expansion capacity (Hamel & Prahalad, 1995, p. 233-241). They are above all a distinctive and unique factor that marks an organization or activity.

Based on the above concepts, informational competence can be defined as the set of professional, organisational and key competences that may be linked to the profile of an information professional or an activity intensively based on information. This competence can be expressed by the *expertise* in dealing with the informational cycle, with information technologies and informational contexts.

The competences information is mobilized in working situations be one of the requirements of the professional profile necessary to work with the information, regardless of the type of professional or activity. They are competence that premise the business processes, the management processes and the technical processes various parts of the same organization or activity. It would be desirable that informational competences were part of the list of competences of the most varied professionals, activities and organizations.

Information Professional

The professional expression of information, in a certain way, has already established itself as a reference to designate, in a generic way, those who work in the area of science of information, although this expression is also used in other areas, as is the case of communication, in which journalists and reporters are often called information professionals. Access to information or facilitating this access is the objective of these professionals, who serve users, beneficiaries of this access, divided into groups with different characteristics and demands of information.

There are many denominations used to refer to professionals who deal with information: Information Professional, information agent, knowledge professional, knowledge worker, archivist, librarian, museologist, communicator, among others and whose relevant activities are linked to the process of

generation, dissemination, recovery, management, correction and use of information. They are professionals whose denominations in the labor market indicate a strong interaction of skills and technical knowledge and management available in the area of the professionals of the information and other areas, such as Education, Marketing, History, Administration, Economics (web designers, content engineers, information architects, among others).

According to Tarapanoff et al (2002), information scientists are the information mediators, since they act between the digital world and the real capacity of understanding the information receivers, guaranteeing the effective communication and the satisfaction of the informational needs of users of this technology. They are professionals who know how to use information technologies; dominate at least one foreign language (English) and have knowledge of Web and tools for connectivity.

The informational cycle identifies all phases of work with information: determination of information needs, sources, collection, processing (information and communication technologies, as support) / treatment (pencil and paper technology), the use and distribution of information. The informational context is the context in which the informational cycle takes place and influences its flow and its characteristics: technological information, business information, scientific information, etc.

Information professionals' skills profiles can be summarized as follows:

- Domain of more than one language, (English required);
- Communication capacity;
- Good interpersonal relationship;
- Management knowledge, (mainly the strategic management of information for the management of organizations);
- Organizational culture;
- Leadership capacity;
- Ability to filter information by adding value to products and information services;
- Ability to work the abstract (teaching / learning of informational tools, management, planning and research), in which the knowledge of information technologies and communication in the user's perspective, are important, but not determinants;
- Ability to assess, plan, sell and operate local information communication networks in institutions;
- Ability to manage information units and implement information management programs, including the Informatization;
- Ability to search, prepare, summarize and edit information of a scientific and technical nature;
- Ability to direct the writing of scientific journals in publishing companies;
- Ability to organize (acquire, register, retrieve) and distribute the information in its original form or as elaborate products from it.
- Ability to perform social functions; education and mediation. The education relates to literacy in information, and the second, to the animation of collective intelligence. Educating oneself and educating others to the information society is one of the great challenges for information professionals and an important step towards the formation of informational culture in society and, eventually, collective intelligence.

CONCLUSION

For the information to be used, it needs to be valuable. Today people and managers are flooded with information. This requires a process of filtering and conscious management to identify what are the relevant information (strategic information) to support decision making (strategic).

The information adds value when it allows people and managers to identify opportunities (strategic and threats), detecting trends or potential problems. The information adds value to the business, when its analysis / interpretation identifies new ways of doing business, new products and or services (product development or diversification).

The relevant information affects the success of organizations, since they are a strategic resource of high added value, because they allow managers to be constantly attentive / informed about the alternative scenarios (external).

Although there is consensus on the importance of information, the available literature mainly covers the electronic treatment of data and reflection. The effective use of information has not grown in the same proportion as technology, which itself is unable to provide the information that people, and managers need.

Future Research Directions

The study of the role of information in the formulation of the business strategy, the information sciences recognize a gap in the set of theoretical publications with potential for enrichment and exchange of interdisciplinary knowledge among the sciences of Information and other fields of knowledge that focus on the strategy.

The scarcity of works focused on these topics in the information sciences can be an indicator not only of the potentiality of spaces to deepen the interdisciplinarity between information sciences and other areas of knowledge, but also the Potential of niches for unpublished and original works in this area of knowledge.

By identifying the competences of some professionals who exercise any activity in an organization and by comparing these activities and specific competences of so-called information professionals, it is likely that there are some "Coincidences", especially if the studied activity intensively uses the information. These "coincidences" are called informational competencies.

REFERENCES

- Brandão, H. P. (1999). *Gestão baseada nas competências: um estudo sobre competências profissionais na indústria bancária*. Dissertação (Mestrado) – Universidade de Brasília, Brasília. www.mec.gov.br/semtec/educprof
- Cleveland, H. (1982). Information as a resource. *Futurics*, 6(3/4), 1–6.
- Dahlberg, I. (1978). Teoria do conceito. *Ciência da Informação*, 7(2), 101-107.
- Davenport, T. H., & Prusak, L. (1998). *Working Knowledge*. Harvard Business School Press.
- Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance*. Academic Press.

Fleury, A., & Fleury, M. T. L. (2001). *Estratégias empresariais e formação de competências: um quebra-cabeça caleidoscópico da indústria brasileira*. 2. Atlas.

Fleury, M. T. L. (2001). Aprendizagem e gestão do conhecimento. In *Gestão por competências*. São Paulo: Gente.

Gianfaldoni, P. (2002). *Normalisation de l'organisation productive et gestion des compétences*. Disponível em: <http://sceco.univ.aix.fr/cefi/workingpap/9914.pdf>

Grant, R. (1991a). The Resource-based Theory of Competitive Advantage: Implications for Strategy Formulation. *California Management Review*, 33(3), 114–135. doi:10.2307/41166664

Hamel, G., & Prahalad, C. K. (1995). Competindo pelo futuro: estratégias inovadoras para obter o controle do seu setor e criar os mercados de amanhã. Rio de Janeiro: Campus.

Le Coadic, Y.-F. (1997). *La science de l'information* (2nd ed.). Paris: Universitaires de France.

Le Coadic, Y. F. (1998). *Le besoin d'information*. ADBS Editions.

Nordhaug, O. (1998). *Competence specificities in organizations*. International Studies of Management & Organization. http://www.iresco.fr/labos/lsci/rite/publication/textes_en_ligne/actes2.pdf

Penrose, E. (1959). *The Theory Of The Growth Of The Firm*. Basil Blackwell.

Prahalad, C. K., & Hamel, G. (1990). The Core Competence of the Corporation. *Harvard Business Review*, 68(3), 79–91.

Tarapanoff, K. (2002). Funções sociais e oportunidades para profissionais da informação. *DataGrama-Zero – Revista de Ciência da Informação*, 3(5).

Teece, P. e S. (1990). Firm Capabilities, Resources and the Concept of Strategy. *Cahier de Recherche*, 90(9).

Teece, P. e S. (1997). Dynamics Capabilities and Strategic Management. *Strategic Management Journal*, 18(17).

Wernerfelt, B. (1984). A Resource-Based View of the Firm. *Strategic Management Journal*, 5(2), 795–815.

Williamson, O. E. (1975). *Markets and Hierarquies: Analysis and antitrust implications*. The Free Press.

Zarifian, P. (2003). *O modelo da competência: trajetória histórica, desafios atuais e propostas*. São Paulo: Senac, Objetivo competência: por uma nova lógica. [http://perso.wanadoo.fr/philippe.Zarifian /page70.htm](http://perso.wanadoo.fr/philippe.Zarifian/page70.htm)

Zikmund, W. G. (2000). *“Bussiness” Research Methods* (6th ed.). Dryden Press Harcourt College Publishers.

KEY TERMS AND DEFINITIONS

Informational Competence: Can be defined as the set of professional, organizational and key competences that may be linked to the profile of an information professional or an activity intensively based on information. This competence can be expressed by the *expertise* in dealing with the informational cycle, with information technologies and informational contexts.

Key Competence: Is the set of skills and technologies whose brand of authenticity is integration. They represent a perceived value by the customer, a differentiation between competitors, an expansion capacity (Hamel & Prahalad, 1995, p. 233-241). They are above all a distinctive and unique factor that marks an organization or activity.

Organizational Competence: Is the *savoir-faire* of the company in a particular domain, which originates and sustains itself by professional competences, allied to the organizational processes and other Resources, or the ability to combine, mix and integrate resources into products and /or services (Fleury & Fleury, 2001, p. 23). They include the skills of the organization and its processes, skills in techniques and forms of work, service skills and social skills.

Professional Competence: It is related to people or work teams, integrating technical, cognitive, social and affective aspects related to work (Brandão, 1999, p. 28). It comprises knowledge, skills and attitudes or behaviors that will allow the organization's development, in fulfilling its mission (Dutra, 2001);.

APPENDIX

14

Information Professionals ' Skills Profiles Can Be Summarized as Follows

1. Domain of more than one language, (English required);
2. Communication capacity;
3. Good interpersonal relationship;
4. Management knowledge, (mainly the strategic management of information for the management of organizations);
5. Organizational culture;
6. Leadership capacity;
7. Ability to filter information by adding value to products and information services;
8. Ability to work the abstract (teaching/learning of informational tools, management, planning and research), in which the knowledge of information technologies and communication in the user's perspective, are important, but not determinants;
9. Ability to assess, plan, sell and operate Local information communication Networks in institutions;
10. Ability to manage information units and implement information management programs, including the Informatization;
11. Ability to search, prepare, summarize and edit information of a scientific and technical nature;
12. Ability to direct the writing of scientific journals in publishing companies;
13. Ability to organize (acquire, register, retrieve) and distribute the information in its original form or as elaborate products from it.
14. Ability to perform social functions; Education and mediation. The education relates to literacy in information, and the second, to the animation of collective intelligence. Educating oneself and educating others to the information society is one of the great challenges for information professionals and an important step towards the formation of informational culture in society and, eventually, collective intelligence.

A Primer on Q-Method and the Study of Technology

Stéphanie Gauttier

Grenoble Ecole de Management, France

INTRODUCTION

Q-method has been invented in 1935 by William Stephenson in order to capture individuals' operant and subjective points of view on a given topic, based on the use of concurrence theory and a forced distribution matrix. Essentially a qualitative research method, it is often qualified of quali-quantology or mixed method due to the use of a q-factor analysis to examine the data.

Q has been largely overlooked by researchers interested in the study of technology. For instance, in Information Systems (IS) research, i.e. research focusing on digital technology as a socio-technical artefact, only 20 papers can be retrieved (Gauttier et al. 2016; Gauzente, 2013). Yet, Q-studies have been published in major journals in the field (among others: MISQ, I&M, CAIS, JIT, OMEGA), indicating its potential for research related to technology. The fit between Q-method and IS topics has been mentioned several times in the literature (Gauzente, 2013; Thomas and Watson, 2002; Kendall and Kendall, 1993; Dos Santos & Hawk, 1988).

This entry describes Q-method and how it can be applied to IS topics to yield new insights. It provides a description of the current use and potential uses of Q in IS research as well as practical advice on how to set up a Q-study. Areas for future research are outlined. It is targeted towards any researcher interested in capturing attitudes to and experiences of technology, as well as those looking for methods to inform the design of technological artefacts and their evaluation.

The reader will notice that the different elements at play during a Q-study are designated as Q-(study, set, factor analysis, etc). The presence of the Q letter could be omitted, and other words could be used. However, the omnipresence of the letter Q in this text is consistent with the vocabulary used by Q-methodologists in publications. It also serves as a reminder to the reader that in this chapter every notion should be interpreted in regard to the Q approach and intentions, and not to the R logic with which the reader might be more familiar. Indeed, in contrast to positivist approaches and the use of R-methods, Q-method considers the self-referent subjective experience of reality (or here technology). It does not aim at a generalizable objective view of a phenomenon, nor does it look at the relationship between variables composing the phenomenon. Q-method identifies the shared perspectives on a given topic between individuals (Watts and Stenner, 2012).

BACKGROUND

Q-method was developed by the psychologist Stephenson (1935; 1953) as an approach to capture people's operant subjective views of phenomena. Subjectivity is conceptualized as what '*emanates from a particular vantage point*' (Brown, 1993). Operancy refers to the fact that these views drive individuals'

DOI: 10.4018/978-1-7998-3479-3.ch120

behaviors. It is a method suited to the identification of the drivers, barriers, and structuring elements of behaviors and experiences.

Q-method rests on two important pillars. One is theoretical and refers to concurrence theory, the other is methodological and uses Q-sorting procedure and Q-factorial analysis (Gauzente, 2010). First, the concurrence theory posits that meaning is dependent upon context and therefore not given *in abstracto*. The concurrence can be defined as the volume of available statements on a topic and is '*the common coinage of societies large and small, and is designed to cover everything from community gossip and public opinion to the esoteric discussions of scientists and philosophers*' (Brown, 1993). Meanings exist for each individual and vary depending on circumstances, but can also be shared with others, thus making interpersonal communication and interpretation possible.

The first step to conduct a Q-study is to generate these meanings. These constitute the Q-sample.

Then these meanings are ranked onto a Q-sort grid, i.e., respondents rank-order assertions according to the degree with which they represent their subjective view of one topic. The forced ranking distribution means that only a small amount of assertions can be selected as highly positively or negatively representative. The majority of meanings will be neutral. This process forces respondents to choose and structure their point of view. The respondents are designated as the P-sample. The result of the Q-sorting process by the participants is a Q-sort. The participants are invited to comment on their Q-sort and reveal how they interpret the elements of the concurrence as well as their own subjective point of view as revealed through the process of the study. Then, a factor analysis is performed to process the Q-sorts. Instead of individuals, assertions or statements are analyzed. In other words, the correlation matrix relies not on assertions but on individuals. This procedure is called Q-factor analysis. As a result, one identifies Q-factors, which are designated as view, i.e. shared views amongst participants. The views shouldn't be assimilated to groups of people as in typological approaches. The views are not a statistical representation of groups in the general population. Rather, they are shared operant views and interpretations of a topic.

Q-method is used in many different disciplines such as political sciences (on questions linked to evaluation and decision-making, but also to capture the structure of individuals' political views), education (capturing learning experiences), health studies (capturing carers' and patients' experiences), leadership studies, but is currently under-represented in IS research.

FOCUS OF THE ARTICLE

The study of technology, as pursued in IS research, has traditionally adopted a positivist lens. In spite of a turn recognizing the potential of qualitative interpretive approaches grounded in social constructivism, sociomateriality (Scott & Orlikowski, 2013), or phenomenology (Boland, 1986) to study technology, the share of such research in major IS publications is still low compared to positivist studies. One can cite two main barriers to the diffusion of interpretive research in IS. First, interpretive research refers to a complex epistemological positioning, with a plethora of traditions and approaches that the researcher must choose from. These approaches are often described at a philosophical level, and little guidance is available to implement them. Second, interpretive research can be seen as biased by positivists: the process of interpretation seems to depend on the subjective and hidden choices made by the researcher and does not offer the possibility to reproduce the process of analysis. Q-method allows overcoming these barriers (Gauttier, 2017), while identifying what drives and structures individuals' relationships to technology.

1. Epistemological Positioning

Q-method is a qualitative method, which has been positioned as phenomenological and interpretivist.

According to Shinebourne & Adams (2007), Q can be qualified of phenomenological as it seeks meaning *‘through exploring subjective accounts of phenomena from participants’ perspectives, attempting to identify broad categories and common themes and a commitment to a collaborative engagement with participants.’* The classification is controlled by the participants as they rank and give meaning to assertions, while the role of the researcher is to exert discretion during the processing of the data (see 2.3) to ensure the views of the participants are correctly represented in the results. The views are given from the participants’ perspectives, interpreted by the researcher.

2. Carrying Out a Q-Study

Descriptions of the process of a Q-study have been given numerous times in the literature (Brown, 1993; McKeown & Thomas, 2013; van Exel and De Graaf, 2005, etc). Details are given concerning 1) the formation of the concourse and Q-set; 2) the administration of the Q-study; 3) the data analysis.

2.1 From Concourse to Q-set

The concourse can be of three different types:

- 1) Naturalistic, i.e. gathered for the purpose of the study at hand through interaction with individuals during focus-groups or interviews (Brown, 1980). The individuals that helped to generate the concourse can be the participants of the subsequent steps of the study, although it is not mandatory.
- 2) Ready-made, i.e. already constituted in the literature or obtained through a literature review or constituted by a series of experts.
- 3) Hybrid, i.e. the researcher completes a naturalistic concourse with elements found in the literature.

There is no type of concourse superior to another. Rather, the researcher must choose to form the concourse in the way that is the most appropriate for the study at hand.

The meanings included in the concourse can then be expressed in the form of pictures, textual statements, and sounds. One could also consider using smells. The vast majority of Q-studies are performed using textual elements. Pictures are used to interrogate children’s points of views, to study arts, or in human-computer interaction to look at preferred designs (for instance such an approach was implemented by the International Society for the Scientific Study of Subjectivity in 2015 to survey members’ preferences for the new website).

The concourse constitutes the available list of elements on a topic, and as a consequence, it can be rather large. There is no magic number on the number of assertions to keep. Q-studies with as many as 1575 statements - even if broken down into smaller sets to administrate the study (Hilden, 1958) and as few as 14 statements (Manera & Wright, 1981) have been conducted, never mind their relevance. However, it is noteworthy that a bigger amount of statements decreases the chance of having random correlations. At the same time, one should consider the format of the assertion and the capacities of participants: children, patients with serious conditions, elderly, might not be able to rank as many assertions as other populations. Brown (1980) recommends a set of 40 to 50 assertions. Sets of 24 statements have been discussed as potentially not enough to cover the nuances of points of view even though they might be

enough for a given topic (Q-method listserv, 2017). In essence, the researcher is responsible for making sure the participants can express their point of view through the assertions proposed while maximizing the quality of data collection. Indeed, given that the participants have to consider these elements holistically to rank them in the Q-sort grid, it is necessary for the researcher to think of the maximum number of elements that can be kept for the final sorting procedure (the Q-set) so that participants can handle the amount of assertions and rank them. Different strategies can be adopted to ensure the representativeness of the Q-set: one can gather the concourse and transform it into a Q-set with participants through focus groups (naturalistic approach); or one can pilot the Q-set and ask participants whether they think the concourse allows them to represent their point of view, then refine the Q-set in accordance to the answers before administrating the Q-study.

The Q-set must avoid redundancies. As the sorting occurs using a scale from positive to negative, there is no need to keep opposite assertions: if a participant disagrees with the statement “eating ice-cream makes me happier”, (s)he can put it in the negative scale on the grid, eliminating the need for a statement “eating ice-cream doesn’t make me happier”. There is no need to keep statements that carry the same meaning as they would be placed in the same column, putting other statements in higher or lower columns only in an artificial manner and rendering the exercise cumbersome for the participant.

The Q-set might include what is seen as a double-barreled statement in so far as in context and for the participant, it is not double-barreled. For instance, a statement taken from an interview and included in a q-set might seem to be double-barreled, but for the interviewee, it is one statement, one unit of meaning, not two, and therefore can be ranked and interpreted as one statement. Moreover, Brown (Qserv list, 2018a) explains that ‘*The other statements surrounding this particular (double-barreled) statement within Q-sort 1 will provide hints as to the particular connotation of the statement, and it is this that demonstrates the gestaltist principle in Q—that the parts gain meaning in relation to the whole.*’ Each statement takes a particular meaning because of how it relates to others, lifting ambiguity on how it is to be interpreted.

In order to assess the understandability of the Q-set and check for redundancies, it is possible to pilot the Q-set before administrating the studies.

2.2 Administrating Q-Studies

Five elements need to be taken into account when one administrates a Q-study:

- 1) The selection of participants: Q-method is, in essence, a qualitative method as it looks at the subjective. It can be used with a small number of participants (Van Exel & De Graaf, 2005). Single case studies are even possible (Stephenson, 1974), as Q aims at representing the point of view of individuals in an extensive manner. In Q, the selection of participants is purposive (Brown, 1980): they are chosen because of how they relate to the topic and/or question researched, i.e. one might want people with specific experiences or characteristics to participate. For instance, in IS research individuals with experience of a specific technology (Gauttier et al., 2016), or individuals who identify themselves with given forms of technology use (Gauttier & Gauzente, 2018) might be selected. Brown (Q-method listserv, 2018b) suggests designing a p-set matrix. This matrix can include key characteristics of participants pertaining to the object of the study, and characteristics pertaining to participants as individuals. It allows designing a participant selection strategy. For instance, to study the drivers and barriers acceptance of technology, one can consider levels of use so that the first line of the matrix is composed, for instance, of a) heavy user; b) occasional user;

and c) non-user. Then, the researcher might want to integrate considerations of gender, adding another element to the matrix, which takes a 3x2 structure.

Assessing the representativeness or weight of each view identified through the Q-factor analysis is not the objective of Q-method. Such a typological approach would be a misunderstanding of the underlying logic of the method. The views have value as they signal the existence of hindrances and drivers of behaviors, with these hindrances and barriers being operant and not only plausible. Comparing views is thus possible, no matter how many people share the view. (McKeown & Thomas, 2013).

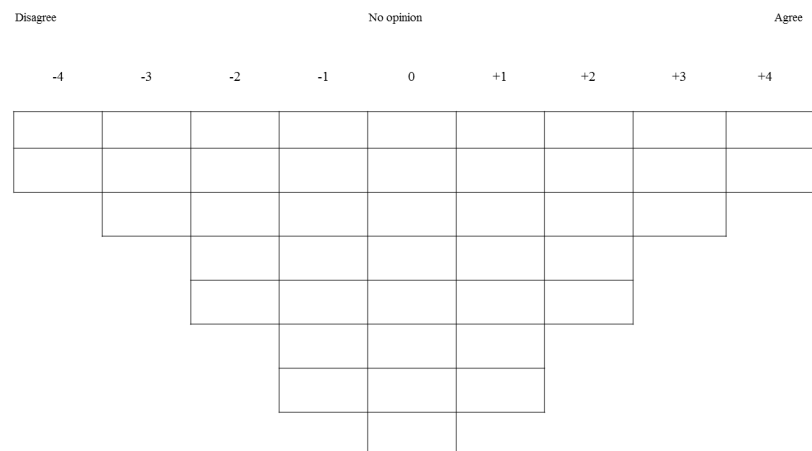
- 2) The conditions of instruction: Depending on the topic of the study, the researcher may opt for a simple design requiring the individual to sort the assertions only once from his/her point of view. Other designs can be implemented. Firstly, one can modify the number of times one proceeds to the sorting so as to see how one's point of view evolves over time, given that some form of intervention is going on. Secondly, one can ask participants to sort the same assertions but looking at various angles of the topic. Thirdly, one can ask participants to proceed to the Q-sort from the point of view of someone else.

For instance, Gauttier and Gauzente (2018) asked participants to proceed to Q-sort 9 times to express their view on mobile technologies overall, different technologies, the point of view participants attributed to their parents, and the point of view participants would attribute to future/children's generations, and then general opinion again.

One could also imagine that instead of having individual Q-sorts, a group could be asked to proceed to Q-sorting.

- 3) The design of the forced distribution matrix: The matrix that will be used to proceed to Q-sorting follows a (quasi) normal distribution, allowing participants to place a few assertions to the extreme, and then more and more, to reach a peak in the middle (see Figure 1). The number of cells in the matrix must match the number of assertions. Recommendations about how to build the scale have been given by Brown (1980, p.200): *'As to the range of the distribution: naturally, the larger the number of statements, the wider the range of available scores should be. As a rule, Q samples smaller than N == 40~can safely utilize a range of +4 to -4; from 40 to 60, a range of + 5 to - 5 is generally employed; beyond 60, +6 to - 6 is not untypical, although there are few occasions for a wider range to be utilized since Q samples exceeding 60 are rarely required; most Q samples contain 40 to 50 items and employ a range of +5 to - 5 with a quasinormal flattened distribution.'*

Figure 1. Example of a forced distribution matrix



- 4) The choice of medium: Data collection can be facilitated online through the use of specific platforms (at the time of writing, WebQ and Qsoftware are used). Some studies have used Skype to conduct the Q-sort, sending the assertions and grid to the participants in advance. It can also be conducted face-to-face. Key criteria for choice are mainly linked to convenience and access to samples. It is however important that the medium allows following the procedure described below.
- 5) The procedure to follow: Firstly, participants are asked to read through the list of assertions entirely, so they have a holistic perspective of it. Then, participants are asked to look at each assertion again and class it in one of three piles (agree, don't agree, no opinion). This helps participants to get acquainted with the Q-set and to proceed to a first sort of ranking, which should help them when they have to produce the detailed ranking demanded in the forced distribution matrix. Q-methodologists do not record this initial rough classification and participants are allowed to modify this classification ranking, for instance, an 'agree' or 'no opinion' as negative in the forced distribution matrix if they want to. The third stage is, therefore, ranking in the grid. Participants can modify their ranking as much as they want and the researcher records only the last version of the grid that satisfies the participants. All assertions must be ranked, and the grid must be respected. The order of the statements in a column is not important.

Sometimes participants have more assertions they disagree with than ones they agree with, meaning that elements ranked under the longest column are not neutral. This is perfectly acceptable as it still means that the statements going to the middle are not the ones the participant focused on and that drives his point of view. Participants can rename the column if they feel they must to convey their idea. Then, a post-sorting interview is conducted where participants have the possibility to explain how they proceeded to the ranking as well as the subjective meaning they give to the statements. Typically, the participants are asked to explain how they choose the statements that are the most extremes and are asked to summarize their point of view.

2.3 Data Analysis

Data is processed through free software. PQmethod has been validated by the Q-methodologists and used in most studies. Ken-Q has been more recently developed and its validation is still in process.

The data is analyzed through a Q-factor analysis process, which is an inverted factor analysis (Stephenson, 1936) in so far as a Q factor analysis looks for what is common between people, instead of what is common between variables.

Table 1. A representation of Q and R matrixes (adapted from Gauzente, 2013)

Q-Matrix			R-Matrix		
Assertion	Individual A	Individual B	Individual	Assertion 1	Assertion 2
1			A		
2			B		
3			C		

Firstly, one needs to proceed to the factor extraction through centroid method (Brown, 1980) or through the Principal Component Analysis method.

Secondly, one rotates the factors, commonly through Varimax. The rotation can be done according to statistical or theoretical motives. Theoretical rotation refers to rotating the factors so as to choose a solution that can be interpreted and explained, even though it is maybe less satisfying statistically. On each factor, defining factors are flagged, based on factors loading.

Then, one obtains the following data: views identified and how the individual Q-sorts load on each factor, a correlation matrix, a table with eigenvalues and variance, a synthetic Q-sort for each factor (each statement with its place on the grid and z-score), distinguishing statements between factors, consensus statements between factors, and arrays of differences. The analysis can lead to the identification of bipolar factors, with significant Q-sorts that load in positive and negative ways.

It is important that the researcher first analyses each view independently of the others, capturing the nuances in each view, before adopting a holistic perspective looking at differences between them.

SOLUTIONS AND RECOMMENDATIONS

Q-method brings results that are of different nature than the ones of R studies. While R studies aim at giving results that are generalizable and representative of a given population, Q gives a representation of the subjective points of views of participants. The non-generalizable character of the results obtained through Q-method is often understood as a limitation, one which is actually common to other qualitative methods. Q-studies are not subject to any form of validation through replication: they capture individuals' points of view as they are, not attempting at verifying their truthful or objective character (Stephenson, 1988). Another limitation can be seen in the presence of subjectivity throughout the process (designing the Q-set, making choices during the factor analysis, and then the subjectivity of the interpretation). However, the role of the researcher's subjectivity in constructing data is a problem common to many methods. Q offers the advantage of allowing to document this subjectivity at different stages: the criteria used to structure the concourse and the steps followed for the factor analysis can be explicitly formulated. The interpretation of qualitative data is therefore open to contradictory analysis as the whole process is traced (Gauzente, 2013). Besides, Q-method allows documenting the researcher's subjectivity alongside the participant's by using the researcher as a participant (Gauttier, 2017).

The interpretation remains subjective but can be verified for its plausible character as they are grounded in mathematical results (the z-score of the statements, the array of difference for statements between factors). Q-studies are easily reproducible as the Q-set is shared with results (Gauzente, 2013; Gauttier & Gauzente, 2016). Q-method allows reaching a deep understanding of individuals' perceptions of a topic as it respects nuances (Gauzente, 2013), and allows to capture complex relationships between points of views (bipolar factors, negative loadings).

With criteria of transparency and reproducibility of both the data collection and data interpretation stages, Q qualifies as a rigorous research method.

At the same time, Q-method is characterized by a process allowing Researchers to be creative to find new ways to shed light on research topics in a qualitative manner. Q can be used to document topics traditionally tackled by studies of technology. In her recent synthesis on Q-method in Information Systems Research, Gauzente (2013) identifies three types of utilization of Q-method:

- Profile identification and fit evaluation
- Identification of structuring trends as an alternative to Delphi technique
- Deep understanding of attitudes, representations, and perceptions.

Over the last five years, a number of papers applying Q-method to IS topics appeared, making use of the creative aspects of Q. They mainly look at attitudes and drivers of perception:

- Study of non-use (Gauttier & Gauzente, 2018)
- Design evaluation (O’Leary et al, 2013)
- Cultural comparison (Gauttier et al., 2016)
- Exploration of TAM and UTAUT (Laden et al., 2018)

Another group of papers comes to address more methodological issues, suggesting that Q can complement our understanding of previous theories, filling in gaps, as well as capture phenomena so far left in the dark:

- Q to study motivations of non-use (Gauttier & Gauzente, 2018)
- Q to perform single-case studies (Gauzente, 2015), i.e. to provide the extensive description of the point of view on a topic, or technology, of one individual.
- Q as a complement to interviews, increasing the transparency of data interpretation in the context of IS study (Gauttier, 2017)

Given the strength of Q-method and its ability to complement traditional research on technology, its use and development are recommended to IS researchers.

FUTURE RESEARCH DIRECTIONS

Q-method can be used to shed light on topics so far explored with different methods. Q-method can also be the object of further methodological research, of interest to IS researchers as well in order to reinforce the place occupied by qualitative research in the field. Four areas are identified:

- *Complementarity between methods suitable for IS studies and Q*: Studies looking at the complementarity between Q-method and R data (Thompson et al., 2013), Q-method and interview data (Gauttier, 2017), have been conducted. One could investigate the complementarity between Q and other methods such as ethnography data, applied to the study of technology.
- *The phenomenological character of Q-method*: A phenomenological turn was called for in IS in 1986 (Boland) and accomplished only to a small extent. Q could be seen as a way to accomplish this task if one could establish in a more detailed manner how Q can be used best to fit the concepts of phenomenology.
- *The use of Q to document the researcher’s subjectivity* can be explored further. This could lead to use the method in combination with other qualitative methods, as a validation process, which could reinforce the acceptability of qualitative research overall in the field of IS.
- Further research could explore topics through different senses, using *sounds, picture, text* to question technology, opening a new way to research the relation to and experience of technological artifacts.

CONCLUSION

This chapter showed that Q-methodology offers a large potential to be explored by IS researchers. The versatility of Q allows designing creative and innovative approaches to data collection, which can be applied to any topic. Through its flexibility, the method can allow to yield new qualitative insights on trends, fits, and perception of technologies which can allow overcoming gaps presented by major IS theories developed through quantitative studies such as the technology acceptance models. Perhaps more importantly, Q can be used to support the generation of high quality qualitative data and, or, strengthen the reliability of the interpretation of other forms of qualitative data, which is much needed in a field of studies where quantitative data remains mostly used, and lead to developing a new stream towards the development and study of technology.

REFERENCES

- Boland, R. J. (1986). Phenomenology: a preferred approach to research on information systems. In *Trends in information systems* (pp. 341-349). North-Holland Publishing Co.
- Brown, S. R. (1972). A Fundamental Incommensurability between Objectivity and Subjectivity. In S.R. Brown & D.J. Brenner (Eds.), *Science, Psychology, and Communication: Essays Honoring William Stephenson*. New York: Teachers College Press.
- Brown, S. R. (1978). The Importance of Factors in Q Methodology: Statistical and Theoretical Considerations. *Operant Subjectivity*, 1(4), 117–124.
- Brown, S. R. (1980). *Political subjectivity: Applications of Q methodology in political science*. Yale University Press.
- Brown, S. R. (1993). A primer on Q methodology. *Operant Subjectivity*, 16(3/4), 91–138.
- Gauttier, S. (2017). Increasing Transparency in Interpretive Research: Q-Method to Objectivize the Researcher's Subjectivity. *AMCIS*.
- Gauttier, S. (2017). Bringing Design Intentions Into The Evaluation Loop: An Invitation To Conduct User Evaluation With Q. In G. Kishmihok, S. Harrison, C. Weber, & J. Karanovic (Eds.), *Crossing Borders in the Comprehensive Investigation of Labour Market*, Book of abstracts. Academic Press.
- Gauttier, S., & Gauzente, C. (2015). Why do certain consumers avoid new media developments? An investigation of three prudent users' subjectivity. *IARIGAI*.
- Gauttier, S., & Gauzente, C. (2016). Are AR shopping services valued the same way throughout Europe? A four-country Q-investigation. *SIM*, 21.
- Gauttier, S., & Gauzente, C., (2017). Exploring the Similarities Between Users and Non-Users of Consumer Mobile Internet Services - Towards a porosity model. *International Journal of Technology and Human Interaction*.
- Gauzente, C. (2013). Une invitation illustrée à utiliser la Q-method dans les Recherches en Systèmes d'Information. *Systèmes d'Information & Management*, 18(2), 69-109.

Hilden, A. H. (1958). Q-sort correlation: Stability and random choice of statements. *Journal of Consulting Psychology*, 22(1), 45–50. doi:10.1037/h0043652 PMID:13513875

McKeown, B., & Thomas, D. B. (2013). *Q methodology* (Vol. 66). Sage publications.

Q-method listserv. (2017). *Question about the minimum number of statements used for a Q-study*. Accessible through <https://listserv.kent.edu/cgi-bin/wa.exe?REPORT&z=4&l=Q-METHOD&L=Q-METHOD&X=2896A51501AD73FAFB&Y>

Q-method listserv. (2018a). *Question about Double Barreled Statements*. Accessible through <https://listserv.kent.edu/cgi-bin/wa.exe?REPORT&z=4&l=Q-METHOD&L=Q-METHOD&X=2896A51501AD73FAFB&Y>

Q-method listserv. (2018b). *Sampling the general public for a Q-study?* Accessible through <https://listserv.kent.edu/cgi-bin/wa.exe?REPORT&z=4&l=Q-METHOD&L=Q-METHOD&X=2896A51501AD73FAFB&Y>

Shinebourne, P., & Adams, M. (2007). Q-methodology as a phenomenological research method. *Existential Analysis*, 18(1), 103.

Stephenson, W. (1936). The inverted factor technique. *British Journal of Psychology*, 8(3), 219–229.

Stephenson, W. (1977). Factors as Operant Subjectivity. *Operant Subjectivity*, 1(1), 3–16.

Stephenson, W. (1988). William James, Niels Bohr, and complementarity: V - Phenomenology of subjectivity. *The Psychological Record*, 38(2), 203–219. doi:10.1007/BF03395016

Thompson, A. W., Dumyahn, S., Prokopy, L. S., Amberg, S., Baumgart-Getz, A., Jackson-Tyree, J., Perry-Hill, R., Reimer, A., Robinson, K., & Mase, A. S. (2013). Comparing random sample Q and R methods for understanding natural resource attitudes. *Field Methods*, 25(1), 25–46. doi:10.1177/1525822X12453516

Van Exel, J., & De Graaf, G. (2005). *Q methodology: A sneak preview*. Academic Press.

Watts, S. (2011). Subjectivity as operant: A conceptual exploration and discussion. *Operant Subjectivity: The International Journal of Q Methodology*, 35(1), 37–47.

Watts, S., & Stenner, P. (2012). *Doing Q methodology: theory, method and interpretation*. Sage.

ADDITIONAL READING

Brown, S. R. (1971). The forced free distinction in Q technique. *Journal of Educational Measurement*, 8(4), 283–287. doi:10.1111/j.1745-3984.1971.tb00939.x

Brown, S. R. (1984). Mis-understandings, Mis-attributions, and Mis-Qs. *Operant Subjectivity*, 8(1), 28–30.

McKeown, B. (1998). Circles: Q Methodology and Hermeneutical Science. *Operant Subjectivity*, 21(3/4), 112–138.

Mingers, J. (2001). Combining IS Research Methods: Towards a Pluralist Methodology. *Information Systems Research*, 12(3), 240–259. doi:10.1287/isre.12.3.240.9709

Myers, M. D. (1997). Qualitative Research in Information Systems. *Management Information Systems Quarterly*, 21(2), 241–242. doi:10.2307/249422

Ramlo, S. (2015). Theoretical Significance in Q Methodology: A Qualitative Approach to a Mixed Method. *Research in the Schools*, 22(1), 68–81.

Stephenson, W. (1977). Factors as Operant Subjectivity. *Operant Subjectivity*, 1(1), 3–16.

Thomas, D. M., & Watson, R. T. (2002). Q-sorting and MIS Research: A Primer. *Communication of the AIS*, 8, 141–156.

Tractinsky, N., & Jarvenpaa, S. L. (1995). Information System Design Decisions in a Global versus Domestic Context. *Management Information Systems Quarterly*, 19(4), 507–534. doi:10.2307/249631

KEY TERMS AND DEFINITIONS

Centroid: A factor extraction method which uses a specific form of communality estimation detailed in Brown (1980).

Concourse: The volume of possible assertions, in any format, on a given topic.

Conditions of Instruction: Designates the perspective that participants are asked to sort the Q-set from.

Flagged on a Factor: Signifies that a Q-sort is representative of this factor view based upon factor loadings (correlations between the sorters and the factors). Other outputs from Q are based only on sorts flagged on a single factor.

Hand Rotation: Sometimes referred to as judgmental rotation or theoretical rotation, this factor rotation is graphical rather than mathematical. This method allows researchers to rotate based on theoretical considerations rather than statistical ones.

PCA: A standard, mathematically unambiguous factor extraction procedure that can be found in standard textbooks on factor analysis.

P-Set: Those who proceed to the Q-sorting.

Q-Sample: The assertions chosen from the concourse to be submitted to participants for the q-sorting procedure.

Q-Sorting: The ranking of the assertions of the q-sample by participants (P-set) on the forced distribution matrix.

Varimax: Common orthogonal factor rotation method where the sum of the variances of the squared loadings (squared correlations between sorters and factors) is maximized.

Modernizing the Academic Library

14

Jennifer Ashley Wright Joe

 <https://orcid.org/0000-0002-6067-0415>

University of Toledo, USA

INTRODUCTION

The digital age fundamentally changed how academic libraries operate. With the advent of electronic resources the job descriptions and duties of librarians expanded to include many of the same roles that IT professionals traditionally play. These roles include dealing with computers and software on a larger scale, as libraries attempt to stay current and relevant by adding computers to their building. How each library rises to meet these new challenges may be different, but there are a few things that remain consistent among them. Academic libraries in the modern age expect a different type of education and experience from their librarians. The administration expects them to deal with virtual materials and virtual resources, which requires different skills and expertise. These libraries have also adapted their mission statements and functions to reflect the change in their resources and personnel. Some of these changes have come slowly over the last two decades, while others have come fast, but they will not be the only changes that libraries will have to endure. Technology is continuing to change at a pace not seen before in history, and patrons are expecting the library to embrace it.

BACKGROUND

Most people would say that the digital age was born with the internet. However digital technologies are much older than that, dating back to computers built in the 1960s. Libraries at this time were using MARC records, which were digitized and are still in use today (Arms, 2012). Though digital databases and e-books were still far in the future, libraries began using this technology early on to improve the quality of the services that they offered their patrons. Libraries used technology to streamline services on both the back end and the front end, from the records that catalogers used to the way that patrons found materials in the library. Card catalogs became obsolete as computing power increased and those records were digitized. The libraries still had tons of physical materials for people to use in their research, but the digital catalog helped make it easier for people to find those materials.

In addition, telephones, not traditionally considered digital technology, helped the library answer questions without patrons having to be in the library. This was a marked change from previous interaction with patrons, because patrons no longer had to be in the library for them to ask questions. Much of the traditional reference interview relies on the body language of the patron and picking up on clues that they do not know they are giving off. Without the visual cues, librarians had to sharpen their ability to hone in on problems. This remote-access librarianship paved the way for other methods of contacting librarians, including e-mail, instant messenger, and chat services. Learning from what they had done with regards to telephone interactions, librarians adapted their reference interactions to serve their patron

DOI: 10.4018/978-1-7998-3479-3.ch121

populations through these methods. While this greatly increased the contact librarians had with patrons, it did decrease the number of people coming into the library, and that was only the start.

Technological advances led to even more changes for the library. When the world wide web and its contemporaries were established, they brought with them a new type of resource, the web page, which librarians had to learn how to vet for accuracy and timeliness (Arms, 2012). Though there were many contenders for what would become the internet, the World Wide Web persevered as the accepted software, albeit with many changes over time. These changes included the addition of colors and images, as well as behind-the-scenes information that could be registered within the properties of the web page, but there were still questions about the authenticity of the information being presented on many webpages. Because anyone could build a webpage, this led to misinformation and confusion among library patrons, a problem that still continues to this day. Academic libraries in their research capacity especially struggled with this problem as students began incorporating webpages into their research. Confusing the matter were many valid websites which held information that could not easily be accessed in other ways. Organizations and government departments began releasing content on the internet. Access became easy, but tracing the source became hard. Furthermore, studies at this time showed that, while students knew that the library had a webpage, and had access to the internet, those surveyed felt that the library and the internet were two separate things (D'Esposito, 1999, p. 458), leading to more concern that students were not receiving the best advice when it came to research materials. At the academic library, student-patrons also had to contend with professors who would not let them use online resources for fear of misinformation. This practice of barring that type of resource led to students not fully understanding how to vet sources on the World Wide Web for their own everyday use.

Databases were created to house materials that were both available in print and digitally. This led to database aggregators such as Ebscohost, JStor, and CSA. Libraries began allocating more and more of their budgets to these electronic resources. Database aggregators sold materials in bulk package, making the selection of large amounts of data easy. However, as libraries continued to sign agreements with these aggregators, they found that there was considerable overlap in the various collections, and that dropping parts of these agreements would increase the price of the packages, essentially giving libraries less information for more money (Zimmerman, 2010). However, libraries could not afford to drop out of the digital age entirely; students, now familiar with web searching, began expecting resources to be available to them online, without them ever having to step foot in a library. Digital technologies allowed the universities to offer classes online for the first time, swelling their enrollment numbers. However, academic libraries saw a decline in their patron numbers even as enrollment continued to grow. This juxtaposition of events required a response from the academic library.

ACADEMIC LIBRARIES RESPOND TO THE DIGITAL AGE

Early in the transition, academic libraries realized that they needed to change to continue to stay relevant. Some of these changes were simple, and included adding popular databases as they first began, or offering technology classes for their students who had missed out on those opportunities in high school. However, as public libraries and school libraries increased their technological advancements, academic libraries found themselves having to go ever farther to keep up with the demand of the digital age.

EDUCATION AND EXPERIENCE

When the digital age first began, the Master's in Library Science (MLS) was the accepted degree for librarianship, and had been since 1924 (Kennedy, et al, 2007). Library programs taught cataloging, reference instruction, and how to prepare research guides, among other things. Programs focused primarily on interaction with the patron. Even as technology progressed and these interactions were completed over the telephone or internet, the emphasis was still on patron interaction.

This concept began to shift with the introduction of the Master's in Information Science (MIS) and the Master's Library and Information Science (MLIS). Classes offered in these programs were more technology driven, and included metadata, database search systems, and social networking, the idea being that if librarians could understand the inner workings of the technology they were dealing with, they could better assist their patron population with that technology. In many programs, these classes were offered alongside the traditionally offered programs, and would often be listed as interchangeable in the course catalog. For example, at Indiana University, in the School of Informatics (formerly the School of Library and Information Science), Representation and Organization of Information, a class that focuses on theories of classification and metadata, is offered as an alternative to the traditional cataloging class. Students enrolled in the MLS program at IU also have the opportunity to take MIS classes as electives, while many others choose to earn a dual degree in both programs.

In the present age, now that libraries are more comfortable with technology, any of the three degrees discussed are appropriate for a position in an academic library. Many librarians still hold the traditional Master's in Library Science, including younger librarians, but the content of those degrees has shifted dramatically since the first library science degree was granted. About 78% of accredited programs have some technological requirement for all students, regardless of whether these students pursue library science or information science, with 11 programs (approximately 25%) requiring technology skills before admission (Scripps-Hoekstra, Carroll & Fotis, 2014). There is an overwhelming idea that even if a future librarian is not interested in technology, there are some basic skills they will need anyway to perform certain job functions. This has proven to be very helpful in preparing them for the changes in the materials and resources that the academic library uses.

Job-listings also changed with the increase in technological demands. Libraries began recruiting librarians that had experience with electronic resources, database management, and other technological experience. The number one change in job-listings is the advent of the new position, electronic resources manager, or electronic resources librarian. Whatever the permutation of the title, these jobs require intimate knowledge of virtual resources and virtual materials on a scale that had not been seen before. Equally important to the familiarity with the resources was a familiarity with vendor language and licensing agreements that differed from the traditional print publishers. Many of the licensing agreements contain carryover language from print resources (Masango, 2005), but these agreements can be complex and daunting.

VIRTUAL MATERIALS AND VIRTUAL RESOURCES

Virtual materials or electronic resources (e-resources) are now the backbone of many academic library collections. These e-resources include databases teeming with journal articles, conference papers, and book reviews, as well as e-books through various systems, and mp3s and streaming video for media collections. While the process of collecting these materials has grown increasingly complicated over the

last two decades, the importance of electronic resource managers (ERMs) and management software was relatively unknown for much of that time. While many smaller universities struggled to catch up by hiring ERMs in the last decade, other universities were pioneers in this field. Many of them designed their own software to catalog and maintain electronic collections at a time when few alternatives existed. These systems became forerunners of the electronic resource management systems that many libraries use today, but libraries are still finding that it is occasionally beneficial to create systems for themselves, because electronic resource management does not have a one-size-fits-all solution. Complicating matters is the fact that these solutions sometimes come late in the process. With virtual materials now taking up large amounts of the collection, both in terms of number of items and amount of money in the budget, special attention needs to be paid to the collection process.

Some materials, however, will not be digitized. There may be concerns with the process of digitizing them, or with the quality that will result from digitizing. Some digitization projects will be rejected due to copyright concerns. In these cases, there will have to be extra effort from the library to make sure their patrons are aware of what is available in the library, so that these materials are used. Still other academic libraries will have special collections, such as historical items or rare books that would not fit into the digitization model. These libraries are still a vital part of the library system at a university, and can be a great draw for researchers from around the globe. Marketing these physical materials alongside digital materials is vital to their continued existence. Otherwise, the materials will be deaccessioned from non-use.

The other option libraries have in the face of unused materials is to place them in archives. This is not always feasible for libraries due to budget constraints, but it is an option to consider when faced with unique materials. In that case, a dark archive would be the best solution. In addition to keeping these materials in your possession, it ensures that there will be an existing print copy of the material available to the world if the digitized version is lost. This is not only important for the institution holding the last remaining physical copy of an item, but for all libraries that would like to have continued access to items. It is, however, important to exhaust all efforts to find an extant digital copy of materials before embarking on a digitization project at the university level, because it can be costly, both in money and reputation, if mistakes are made.

There are special copyright rules associated with making copies and digitizing materials with the intent to archive them, and this can sometimes circumvent traditional copyright problems with regards to digitalization. Earnest discussion of this in law journals began with the Authors Guild, Inc. v. Google, Inc. case in 2013, in which Authors Guild, Inc. challenged Google, Inc.'s legal right to digitize copyrighted materials, on the grounds that Google, Inc. would be profiting from the digitization. The judge in the case ruled in Google, Inc.'s favor, citing that the in-depth indexing that was the direct result of digitization was substantively transformative, because it would help researchers be able to better find books (Authors Guild, Inc. v. Google, Inc., 2013). A similar case involving HathiTrust ended with the same results. This opens the door for more digitization projects in the future, though it does not negate the time and money required for such an endeavor.

Also of particular interest in libraries when it comes to electronic resources is the cost of materials. The cost of online databases has risen steadily in the last decade and will continue to do so unless suitable alternatives are found (Zimerman, 2010). While many libraries are ready to give in to the cost and adjust their subscriptions and purchases based on it, there are alternatives already available. One such alternative is the open access journal. The Directory of Open Access Journals lists over nine thousand journals in eighteen categories, each housing scores of smaller disciplines. Not only are these journals as good as their for-pay counterparts, but often they are preferred by researchers in their fields. In addition,

materials in open access journals can be supplemented by online repositories sponsored by universities. These repositories hold pre-prints of articles, as well as other materials, depending on the legal agreements the universities can make with the publisher. Many of them also house work that undergraduates and graduate students have done during their time at the university, making for a more well-rounded research body. Not only can students see what is being pursued by researchers in their field, but they can also see what work their fellow students have been doing, encouraging collaborative work within disciplines and without.

MISSION STATEMENTS AND LIBRARY FUNCTIONS

The rise of digital material required a response not only from the librarians out in the field doing work with patrons and acquiring and cataloging materials in technical services, but also from the administration governing these librarians. It has long been accepted that universities would have mission statements at both the university level and at the department level, which would be organized in tandem so that the university was working toward one united goal. As their roles on the university campus changed, academic libraries needed coherent mission statements that encompassed all the work that they were doing, not just their role as repositories for the physical book. Mission statements began including language about the technology, itself. For example, in Western Kentucky University's mission statement, the fourth tenant is, "Maximize digital technologies and develop networked resources that enhance learning (WKU Library Advisory Council, 2013)." These new mission statements do not just deal with library technology as a burden that has to be dealt with, but as a tool for improving library resources and service. Activities that go along with these mission statements include building information commons, where students can gather and partake of information through a variety of means, including print and digital. They often also have televisions tuned to news headlines and group study areas for students to work collaboratively.

This reflects the changing library functions. Libraries are now a collaborative area filled with students working in groups. In a way, this is just the library catching up to what has been done in the classroom for decades. Group projects had often been assigned to students in years past, and these projects were often carried out in the computer lab, or shut away somewhere in a group study room. However, libraries can, with technology, now embrace these group projects and make it easier for students to work. Now the library is a place where that work can be done easily, with access to all of the materials the group would need. Though there are librarians on duty to help students with questions, the stereotype of the shushing librarian has been shattered. This was a change necessitated by falling patron counts; as students began accessing materials on-line, libraries had to find other reasons for students to come into the library. One of the library's best resources is the space that they provide for students to work, and this was a resource that was underused because the space had not adapted to fit changing work demands.

Changes in how librarians view resources, and what constitutes a resource, have also facilitated new growth in the library. Both academic and public libraries have makers' spaces in which patrons can experience new technology that they might not be able to access or afford outside of the library setting. The tools that constitute makers' spaces can be either digital or physical in nature. Examples include video production software, laser cutters, and builder kits. Depending on the patron population, any or all of these might be available at the library, and are usually tailored to meet the demands of certain majors. Though maintenance and materials can be costly, these types of tools have become more resources for the library to offer.

TECHNOLOGICAL PARTNERSHIPS

In a quest to stay relevant and form connections within the campus community, academic libraries have forged technological partnerships with other departments on campus. Some of these partnerships are necessary; today's libraries have a great deal of technology, including computers, printers, televisions, etc., and they often need their own dedicated IT staff to troubleshoot these devices. Furthermore, because the library is a place of study and work, many students bring their own technology into the library. To serve their patrons better, some library IT personnel also answer questions from patrons about devices and software, solving problems much quicker than librarians typically would.

Not all library-technology partnerships are strictly by necessity, however. Some partnerships are forged because libraries have space, where stacks once stood. When materials are digitized and their physical volumes discarded or archived, libraries can open up the space to other academic departments. Other partnerships are created because the library has a mission to embrace all forms of learning and information gathering. These partnerships can take a variety of forms, including makerspaces, virtual reality labs, and listening libraries. The aforementioned makerspaces increasingly popular in libraries even without official partnerships, though they are sometimes sponsored by an associated academic departments to offset costs.

Regardless of their exact nature, these partnerships are mutually beneficial, because they provide a space for the programs involved and in turn bring more people into a library. In an era where statistics are important evidence of worth, the more people counted going through the library doors, the better. Furthermore, these and other types of partnerships foster goodwill in the campus community and provide a moderate amount of publicity for the library (Wainwright & Davidson, 2017). Some innovations, such as learning and teaching labs, get even more out of the partnership, because they can take advantage of the materials the library already provides.

SOCIAL MEDIA

As a digital presence becomes essential for the continued stability of all companies and institutions, academic libraries have also begun embracing social media, whether as an entity on their own or as part of the larger college or university. Social media has many definitions in the literature, but can be broadly defined as, "the means of interactions among people in which they create, share and exchange information and ideas in virtual online communities and networks (Fang, et. al., 2014, p.336)." Most universities have some policy on who can use social media networks on behalf of the university, and the library's ability to control their own online presence would be affected by those policies.

However, an online presence for marketing purposes is not the only reason libraries should embrace social media. Social media can be an effective tool in the classroom for information literacy. Librarians who teach information literacy classes have embraced Twitter, Facebook, blogs, and more in an effort to foster communication between students and between students and their instructors. Even one-shot library classes can use social media for quick communications during the course of the class, through the use of hashtags on Twitter, or polls on interactive websites, or discussion boards on the various learning management systems.

There is also a third reason for academic libraries to embrace social media. Social media networks have become powerful tools for information exchange in a brief amount of time. Because of this, the general public is less educated about how to vet these sources. Academic libraries should take it upon

themselves to see social media as an information literacy topic, to be taught in classrooms and discussed in presentations and publications by the library. Through these methods, students will become better acquainted with this use of social media networks and the potential problems of taking information at face value.

FURTHER RESEARCH DIRECTIONS

The challenges that libraries face today could potentially sound the death knell for smaller academic libraries. Students no longer believe that libraries have the monopoly on research materials, and they are right. Google scholar has become a serious challenge to libraries despite a pushback against it from librarians, because it is attached to a name that nearly everyone knows – Google. Most students, especially undergraduates, can find the materials they need for their class projects from the comfort of their own home computer. Part of the way to combat this problem is to embrace the supposed enemy. Google scholar can be a very effective tool when used properly, and if a student has the system set up correctly, they can even use Google Scholars' search features to search their own libraries. This search feature has been immensely popular with students, but other databases are still important to the research process and may have search interfaces more suited to the search topic at hand. Libraries have to become creative when trying to draw in students to their libraries and to use their resources.

This can be accomplished through a strong connection to the faculty. Students see faculty every day and are more likely to take their advice when it comes to research and their grades. Set up library instruction sessions for classes, and be sure the students know what's in it for them when they attend. Explain to them the difference between searching in Google and searching in a database, and how the latter can actually be easier and quicker to do once they know how to do it. Running through search questions in the various databases can demonstrate to them the different types of results that they can get. Furthermore, it will show them that it can be difficult to retrieve materials when one is not logged in to a university database.

Other suggestions for making a library more inviting and adaptable to the changing digital world include creating collaborative learning spaces that integrate technology with the learning experience, adding outlets and power stations to facilitate the use of laptops, tablets, and other portable electronic devices. Information commons like the one discussed here are an excellent way to facilitate this change and are very popular subjects for funding and grants at this time. They provide a seamless entrance for the library into the world of technology and provide a valuable resource for the students. Furthermore, if a campus has a technology center that is not part of the university library system, an information commons can artfully redirect this traffic back to the library. This way, students will have academic librarians on hand to answer questions when or if they arise, making the research process easier and encouraging them to continue their studies.

The last advice for the academic library in the digital age is to consider changing how success is measured. A library whose success is measured by the number of people walking through their physical gates and using their brick-and-mortar building is going to be disappointed by the numbers that they see. Universities are more flexible with their class models, offering more classes online, and therefore many people may never step foot on the university campus, much less in the university library. Rather than letting these students languish, struggling on their own to accomplish research while their counterparts on campus have all the help they could want, university libraries should be working with advisors and classes to make sure that they are also reaching out to these students. There are a variety of ways that

this can be done, including web tutorials, research guides, and videotaped presentations on library resources that can be embedded in the software that the university uses for their online classes. The tools to reach these students are available, and helping them should count in statistics in a meaningful way. All ways that the university resources are accessed should be given equal weight, because the university is still providing a valuable service, whether librarians are conversing with patrons through email, on the telephone, or in face-to-face interactions. Even keeping track of patrons accessing virtual resources through OpenURL can be an eye-opener to the way that the academic library is helping their students (Wright, 2014).

CONCLUSION

Today's library exists in a very different world. Much of the library's content is digital, and can be accessed online. Some materials, however, cannot be digitized at this time because of technical or copyright concerns. Identifying these materials and integrating them into the curriculum is important for their continued value and to increase patron counts in libraries. Equally important to libraries is the constant innovation of new resources and outreach methods to serve the patron population. Changing how they measure success will also help libraries justify and evaluate programs, because it will highlight the way that their services have changed over the years. For the most part, libraries are already keeping track of statistics in many and varied ways, and acknowledging that these changes exist. However, they are not doing enough to put more emphasis on those reference interactions and that resource usage, when they should be. By embracing the digital age, libraries will survive and thrive far into the future.

REFERENCES

- Arms, W. Y. (2012). The 1990s: The formative years of digital libraries. *Library Hi Tech*, 30(4), 579–591. doi:10.1108/07378831211285068
- Authors Guild, Inc. v. Google Inc., 954 F.Supp.2d 282 (S.D.N.Y. 2013)
- D'Esposito, J. E., & Gardner, R. M. (1999). University students' perceptions of the Internet: An exploratory study. *Journal of Academic Librarianship*, 25(6), 456–461. doi:10.1016/S0099-1333(99)00078-6
- Fang, J., Wen, C., & Prybutok, V. (2014). An Assessment of Equivalence between Paper and Social Media Surveys: The Role of Social Desirability and Satisficing. *Computers in Human Behavior*, 30, 335–343. doi:10.1016/j.chb.2013.09.019
- Kennedy, K., Russell Gonzalez, S., & Cenzer, P. (2007). Student Perspectives on Library School Degrees and the Hiring Process. *Journal of Education for Library and Information Science*, 48(4), 284–293.
- Masango, C. A. (2005). Digital licence agreements and their effects on acquisitions and academic library users. *South African Journal of Library and Information Science*, 71(2), 127–135.
- Scripps-Hoekstra, L., Carroll, M., & Fotis, T. (2014). Technology Competency Requirements of ALA-Accredited Library Science Programs: An Updated Analysis. *Journal of Education for Library and Information Science*, 55(1), 40–54.

Wainwright, A., & Davidson, C. (2017). Academic Libraries and Non-Academic Departments: A Survey and Case Studies on Liaising Outside the Box. *Collaborative Librarianship*, 9(2), 117–134.

WKU Library Leadership Council. (2013). *Dean's Office – Library Mission*. Retrieved from <https://www.wku.edu/library/deansoffice/mission.php>

Wright, J. (2014). Considerations of a Digital Age: The Hows and Whys of Electronic Resource Management from a Collection Development Perspective. In N. Patra, B. Kumar, & A. Pani (Eds.), *Progressive Trends in Electronic Resource Management in Libraries* (pp. 17–30). doi:10.4018/978-1-4666-4761-9.ch002

Zimmerman, M. (2010). Periodicals: Print or electronic? *New Library World*, 111(9/10), 426–433. doi:10.1108/03074801011089341

ADDITIONAL READING

Becker, B. W. (2015). Advances in Technology and Library Space. *Behavioral & Social Sciences Librarian*, 34(1), 41–44. doi:10.1080/01639269.2015.996496

Chamberlain, E. (2012). Investigating Faster Techniques for Digitization and Print-on-Demand. *New Review of Academic Librarianship*, 18(1), 57–71. doi:10.1080/13614533.2012.660769

Hazen, D. (2011). Lost in the Cloud: Research Library Collections and Community in the Digital Age. *Library Resources & Technical Services*, 55(4), 195–204. doi:10.5860/lrts.55n4.195

Knight, J. (2017). Academic librarians as change champions: A framework for managing change. *Library Management*, 38(6/7), 294–301. doi:10.1108/LM-03-2017-0031

LaGuardia, C. (2012). Library Instruction in the Digital Age. *Journal of Library Administration*, 52(6/7), 601–608. doi:10.1080/01930826.2012.707956

Lippincott, J. (2015). Libraries and the Digital University. *College & Research Libraries*, 76(3), 283–295. doi:10.5860/crl.76.3.283

Mwaniki, P. (2018). Envisioning the future role of librarians: Skills, services and information resources. *Library Management*, 39(1), 2–11. doi:10.1108/LM-01-2017-0001

Roland, L., & Bawden, D. (2012). The Future of History: Investigating the Preservation of Information in the Digital Age. *Library & Information History*, 28(3), 220–236. doi:10.1179/1758348912Z.00000000017

Zabel, D., Shank, J. D., & Bell, S. (2011). Blended Librarianship: [Re]Envisioning the Role of Librarian as Educator in the Digital Information Age. *Reference and User Services Quarterly*, 51(2), 105–110. doi:10.5860/rusq.51n2.105

KEY TERMS AND DEFINITIONS

Academic Libraries: Libraries affiliated with universities or colleges.

Digital Age: From about 1980 on, coincidence with the advent of the world wide web.

Digitization: Scanning or otherwise manipulating print materials into an electronic format.

Electronic Resources: Library materials available in an electronic format.

Mission Statement: Unified theory, plan, or goal of an organization or subsection of the organization.

Social Media: Websites and applications allowing users to network and share information.

Prospects and Challenges of Web 3.0 Technologies Application in the Provision of Library Services

Promise Ifeoma Ilo

University Library, Pamo University of Medical Sciences, Port Harcourt, Nigeria

Christopher Nkiko

Elizade University, Nigeria

Cyprian Ifeanyi Ugwu

University of Nigeria, Nsukka, Nigeria & University of South Africa, South Africa

Justina Ngozi Ekere

University of Nigeria, Nsukka, Nigeria

Roland Izuagbe

 <https://orcid.org/0000-0002-1697-0242>

Covenant University, Nigeria

Michael O. Fagbohun

 <https://orcid.org/0000-0002-3114-3333>

Covenant University, Nigeria

INTRODUCTION

Information and Communication Technologies (ICTs) have continued to redefine and transform the way and manner library and information services are conceived, perceived, offered and used. While it was a subject of debate that the attendant effects of the emergence of the internet and web technologies would mean the devaluation of library and information services and perhaps the demise of the profession, events shaping the information sector have proved that like technology, the field of librarianship is adaptive and evolving. Definitions, concepts, theories and practices once considered “core” or “modern” or “of international best practice” have become obsolete, inadequate and unacceptable due the various paradigms swings that have greeted the profession overtime (Izuagbe, 2017). This further strengthens the argument that the library profession is moving from “building collections” and owning them forever to “renting access” and having ‘connections’ (Hielmcronne, Maiello, Bainton & Bonnet, 2012).

The wide acceptance and adoption of technological innovations in the 21st century libraries is placing emphasis on the application of web technologies due to speed and accuracy (Govanakoppa & Kumara, 2014). Accordingly, libraries that are currently blazing the digital trail are striving towards web services. A web service is a modern application of computing technology aimed at the facilitation of exchange of communication between or among interconnected computers that is internet-based (Amrohi, Garg & Chauhan, 2015). While this service best describe the principles of web 1.0 and 2.0 respectively, it omits

DOI: 10.4018/978-1-7998-3479-3.ch122

the unification of these technologies and users for an all-inclusive problem-solving and knowledge-creation efforts which web 3.0 entails. It is on this note the article sets out to examine the prospects and challenges of web 3.0 application in the provision of libraries services. Specifically, the paper gives insights into the principles and features of the technology, explore areas of applicability relative to library services, highlight the prospects of the technology as well as its challenges within the precincts of library services.

BACKGROUND

The sophistication and rapid dynamism that characterize the ICT have necessitated the need to leverage on latest discoveries in searching, tracking, accessing, disseminating and sharing from the global information network as well as connecting with people and communities for mutual benefits. The World Wide Web has made possible the instant linkage of information resources and people on the cyberspace. It began with the advent of web 1.0. This is a system of interlinked, hypertext documents through the internet. It was a read only platform having focus on one way communication. Users were allowed to search for information and read same with little interactivity and contribution to content. It was aimed at providing online presence for requisite information through websites (Kadyan & Singroha, 2014).

The improvement over this phase resulted in the emergence of web 2.0 in 2004. This version foisted a truly virtual global community where users could share, interact, and collaborate with others on the social media platform. This gave birth to networking sites such as Facebook, YouTube, Blogs, Wikis, Flickr, Twitter and WhatsApp. As a result, the technology has been described as a veritable tool for internet marketing of product and services (Ekwueme & Ilo, 2019). Following in the rapidity of the heightened developments in the programming world, John Markoff mooted on the idea of Web 3.0 in 2006 (Morris, 2011). A web 3.0 is a centric delivery pattern in which relevant data are provided to end users on different platforms and devices in a most usable manner. It has the ability to aggregate information from multiple sources and establishing semantic relationship between all available contents to ensure seamless accessibility, searchability, availability, and usability. It allows interoperability with other devices such as tablets, smartphones, and desktop computers, iPads. It also supports offline use or consumption of downloaded data, therefore making it possible for the exploitation of information in a low bandwidth conditions.

This is made possible by the use of web cookie and client side storage facility present on client device. Kadyan and Singroha (2014) summarized the features of web 3.0 as constituting: intelligence, interoperability, personalization, and virtualization. Web 4.0 also known as symbiotic web aims at interaction between human and machines. It is an open, linked, and intelligent web. It is the basis for the Internet Of Things where smart appliances connected to the internet will be able to perform tasks without human involvement based on the use of radio frequency identification tags (RFID) (Kang & Yong, 2007; Kodyan & Singroh, 2014; Letts, 2018)

The need for the application of web 3.0 technologies in the provision of library services is premised on the ever increasing effect of emerging technologies on education and the growing demand on the capacity of librarians as well as the need for higher strategies in meeting the sophisticated expectations of the ICT-savvy clientele of the 21st century. The present library patrons are desirous of federated library databases that collocate relevant contents from multifaceted sources and facilitate interactive information exchange, scalability, and collaboration on the web. The information seeking behaviour of an average library patron has been revolutionalized by the limitless scope of the Internet and the World Wide Web

in information retrieval. The usual first point of call for information seekers in the twenty-first century is the Internet (Rao & Babu, 2001; Coffman, 2012; Purdine, 2012 & Nkiko, 2014).

FOCUS OF THE ARTICLE

Library 3.0: Application of Web 3.0 in the Provision of Library Services

The usefulness of web 3.0 technologies has become a critical factor that would drive the applications of these emerging technologies to many areas of life. There is a growing interest in the use of these technologies to provide library services. Library 3.0 represents a transformation from Library 2.0 to Library 3.0 with the application of Web 1.0 and Web 2.0 elements. In Library 3.0, library professionals are expected to provide services to library users based on web 3.0 technologies.

Library 3.0 refers to libraries using emerging technologies to facilitate the sharing of use-generated content as well as collaboration between libraries and their clientele. Semantic webs include cloud computing, mobile devices and established tools like federated search systems to facilitate the development and sharing of use-generated content, collaboration between librarians and users, and to promote and make library collections widely accessible, searchable and usable (Belling, Rhodes, Smith, Thomson, & Thorn, 2011). Library 3.0 can make library collections readily available to users irrespective of their physical location. Some of the key elements of Library 3.0 are speed, accuracy, precision and systematic organization, (Chauhan, 2009).

Whereas the origin of the concept of Library 3.0 is controversial, its benefits are inexhaustible. Firstly, it seeks to make all web contents seamlessly accessible, searchable, available and usable (Chauhan, 2009). By so doing, the right information will definitely get to the right user at the right time. Secondly, in the Library 3.0 environment, library users can work together with the Librarians as co-creator of contents and services. Further, it enables library users to make annotations and carefully compare books and other information sources from a variety of useful perspectives. Thirdly, Library 3.0 It promotes dual interaction between computers, OPACs and libraries (Nesta & Mi, 2011).

Fourthly, Library 3.0 is likely to create a network of libraries with distributed collections, multiple interfaces as well as traditional and non-traditional forms of scholarly communication (Libner, 2003). This is premised on the fact that institutions with Library 3.0 will hold large collections in all forms and formats including web pages which will all be accessible anywhere. Finally, Library 3.0 seeks to provide full-text information with a fully integrated deep search environment and provide ways library services can be offered to required services and collections that will exceed patrons' expectations (Bhat-tacharya, 2016).

Principles of Library 3.0

Library 3.0 has some basic principles. Though Library 3.0 principles were scarcely treated in library literature, Saw and Todd (2007) were able to propose five basic principles of Library 3.0. First, the library is intelligent. This means that Library 3.0 uses artificial intelligence (AI) systems to offer intelligent services to library users. As an aspect of AI, intelligent library within the precinct of higher education facilitates personalized leaning, learning analytics and intelligent tutors (Cox, Pinfield & Rutter, 2018). The authors noted further that intelligent library system builds and develops collections and services for

an intelligent agent rather than a human audience. With the aid of intelligent library systems, it is easier to make libraries, not only interactive, but more user-friendly (Dent, 2007).

Second, the library is organized. This means that Library 3.0 model is designed to describe or organize web resources in a manner that users can access them easily. It emphasizes the library coming to the patrons in a proactive and interactive manner, leading them rather than following, innovate, rather than copying (Nesta & Mi, 2011). Third, principle of web 3.0 the library is a federated network of information pathways. Library 3.0 is designed to promote federated searching from a federated list of web resources facilitated by cloud computing systems (Belling et al, 2011). Fourth, the library is apomediated. This explains users sidelining conventional information gatekeepers through the use of technological platforms (Eysenbach, 2008). Whereas social media is helpful in this regard (O'connor, 2010), librarians can apomediate by either providing clues necessary for helping users locate the right information (Sapp & Van Epps, 2006). Finally, the library is MyLibrary. Library 3.0 is designed to provide personalized library services to the users. To achieve this, librarians and users employ Library 3.0 tools to create personal and professional profiles that help to provide user- centred library services, or help to fit library services into the lives of the users (Cohen, Ferreira, Horne, Kibbee, Mistlebauer & Adam, 2010).

Features of Library 3.0

Library 3.0 has some important features. One of the features of Library 3.0 is Web OPAC. This is a library catalogue on the Web or internet. It has a Uniform Resource Locator (URL) which users can connect to and search the required or needed information. Web OPAC has come to replace the card catalogue and has searchable functions for borrowing and automatic reservation. Library users can also request for any kind of information such as contact details, email addresses, new arrivals as well as borrowing rights and privileges. Another feature of library 3.0 is that it makes use of ontologies. In the library context, ontologies are used to provide annotation to web contents. Ontologies are used to show how information is organized. Finally, Library 3.0 enables the creation of personal contents known as ubiquitous contents which can be stored on the web.

Library 3.0 and Virtual Reference Services

Librarians are gradually utilizing the potentials of technological innovations to serve users in remote locations. Bhattacharya (2016) citing Linda Berube (2003) defines virtual reference as a network of expertise, intermediation and resources placed at the disposal of users who need answers in an online environment. This author further state that virtual reference ranges from personal or telephone assistance given to users to developing contents that can be accessed via smart phones and other mobile devices. The author concludes with the following objectives of virtual reference in the Library 3.0 era:

- i. To provide individual assistance and instruction.
- ii. To provide and maintain an appropriate collection of reference resources, both print and electronic
- iii. To assist users with locating the best sources of information.
- iv. To help in marketing reference resources and services
- v. To help in referral process, forward the enquiry or provide the users with live links to authoritative websites.
- vi. To educate users concerning resources and research techniques in order to help them to become information literate.

Reference service based on Web 3.0 environment means new roles for librarians. Some authors have agreed that librarians would function in this new environment as virtual librarians (Alenezi, 2012; Pinto & Manso, 2012). In specific terms, librarians should consider using the library homepage as a reference tool or as a tool for the provision of online reference interview. Furthermore, the authors suggests that librarians should begin to use desktop videoconferencing to provide reference service.

Semantic Web and Library Services

In this digital age, the first source of information for students, researchers, scholars and faculty members is the World Wide Web (WWW). The problem with the www is that its documents or information sources are usually in the form of Uniform Resource Identifiers (URI) and the information in them is not machine readable. However, the semantic web has emerged to address this issue. The semantic web provides a common framework for sharing of data and information in an organization such as a university library. This web consists of both structured and unstructured documents. The semantic web helps to convert the unstructured web documents into web data for global knowledge exchange.

The semantic web is an extension of the current web in which information is given well-defined meaning, enabling computers and the people to work in cooperation (Berners-Lee, 1999). From this definition, the semantic web accomplishes two things; providing machine readable description of web documents and giving meaning to the content on the web documents. When these two things are accomplished, all the machines linked to the semantic web will be able to gather and analyze information automatically in a manner similar to that of human beings. The resultant effect of this is that more relevant and meaningful information will be retrieved. These technologies include web ontology language, Notation 3 (N3), Resource Description Framework (RDF), SPARQL, which is a RDF query language, Turtle (RDF Triple language) and RDF schema (RDFs). Apart from these technologies and standards, semantic web can be realized through artificial intelligence. When this happens, the computer systems will be able to make decisions like a human expert and not by relying on computer algorithms and procedures for this purpose.

In the library, semantic web can be used to develop library portals. The implementation of library portals with semantic web services will definitely help to fulfill the vision of the library. By this implementation, the large collections of the library are annotated semantically by using various software technologies, or this implementation involves annotating the web content information of library resources and providing a semantic description of this content so that it can be machine readable. Formal ontologies or ontology schema are used to provide more flexibility in the semantic description of this web content information. The necessary condition for the functionality of library portal with semantic web services is that the portal's user interface must help the user to search, retrieve and access information from multiple sources.

Furthermore, other library operations that can be made available to semantic web applications are library catalogues, cataloguing of digital resources, collection development policies and reference services. The library catalogue has structured content which makes it have the characteristics of semantic web. For this purpose, the semantic web focuses on the framework for describing information sources and making them available for public use. This, of course, will lead to the development of semantic catalogue that would have ontologies, authority verification and MARC based schema. The semantic web application in collection development begins with ensuring that the collections developed have the characteristics of semantic web.

IFLA guidelines for information services and other resources on digital reference (Lipow, 2005) can provide an acceptable framework for the implementation of semantic web reference service plans.

Furthermore, the semantic web resources are expected to be used by students, faculty members and researchers. Just the same way circulation policy protect library holdings and promote their use, a semantic web resources use policy should be in place for the same reason. That is, this policy should promote web resources use and protect these resources as well. The policy should, as a matter of necessity, provide users with access procedures, a unique identification number and knowledge of their borrowing privileges.

It is important to point out here that the most commonly used semantic web software technology is the Resources Description Framework (RDF). The reason is that it has several applications. Some of them are Friend of a Friend (FOAF), which is used to describe personal information and describe people including the links between them; Simple Knowledge Organization System (SKOS), used to improve recall and precision ratio and to provide searching options or features for users; DBpedia that provides interface that users can exploit to query relationships and properties associated with resources and to provide links to related datasets; and Resource Description Framework for attributes (RDFa) used to annotate web pages.

Cloud Computing and Library Services

This is one of the Library 3.0 applications that are increasing gradually in popularity on daily basis. Simply put, cloud computing refers to the use of internet and central servers that are remotely located to maintain data and other applications as opposed to maintaining data using personal computers or individual mainframe computers (Kwanya et al., 2012). The advantages of cloud computing for library include lower-cost computers for users, better performance, less IT infrastructure costs, less maintenance costs, unlimited storage capacity and anywhere access to documents (Stinivas & Chary, 2017). Many companies are now providing cloud based digital library services. One of them is Duraspace that has two software now in use, namely: “Dspace” and “Fedora commons”. Of the two, Dspace is widely used by libraries to build repositories aimed at providing access to research output and scholarly content. Bhattacharya (2016) summarized the applications of cloud computing in libraries thus:

- i. OCLC is one of the best instances for making use of cloud computing for sharing libraries’ data. Its world cat is the most popular online union catalogue for searching library data.
- ii. To access any files on the internet, cloud computing presents a number of services such as flicker, Drop box, Juggle Disk, Google Doc, Sky Drive and so on. These services virtually share the files anywhere and anytime without any special software and hardware.
- iii. Cloud computing technology offers great opportunities for libraries to build networks among the library and information service professionals as well as other interested people including information seekers by using social networking tools like “Twitter”, “Facebook”, “Instagram” which play a key role in building community power.
- iv. For automation purpose, ‘Polaris’, ‘Ex-libris’ provide variant cloud based services such as acquisitions, cataloguing, process system, digital contents and also support various standards such as MARC 21, XML, Z 39.50.

Finally, examples of cloud service providers for libraries include OCLC, Ex-Libris, Polaris, Scrib, Discovery service, Google Doc and Google Scholar.

Federated Search and Library Services

Libraries introduced portals so as to be able to bring large amount of information together. A portal can be defined as a gateway or a point for information searching on the web. It is seen as the point where students, researchers and scholars or simply users can start to search for relevant and important information on the web. There are indeed different types of portals such as institutional portals, subject portals and federated search tools (Kumar, Sanaman & Rai, 2008). According to these authors, federated search tools are used for integrated access to the resources that the library pays for and they also allow cross-searching of these resources. It is important to note that authentication is often required. Further, the federated search tools work by using Z39.50 protocol for communication between computers. The Z39.50 helps users to extend information search beyond the online virtual catalogues to searching electronic databases that are remotely located but are authenticated at the point of access.

At its basic level, federated search is simply simultaneous searching of multiple online databases or information sources. This means that through one search query users view search results from multiple information sources in a single integrated list. Federated search enables users to search many OPACs, websites, and subscription and citation databases at the same time. It is pertinent to note that federated searching is an important part of an information portal because it provides a single interface to diverse or multiple information sources. For example, once a student enters his or her search query in the search box of the information portal, the query or search string is sent to each information source in the information portal or federated search list. Based on the search query, the individual information sources will send a list of results to the information portal. At this point, the student can view what has been retrieved from each information source in a single integrated list of search results.

Federated searching is somewhat different from traditional search engines such as Google. By using Google, only information sources that have been indexed by the search engine's crawler technology can be searched and retrieved or accessed. The same applies to other traditional internet search engines which means that documents in databases that are not open to these search engines can neither be searched nor retrieved or accessed due to the limitations in crawler technology. Federated search helps to solve this problem by making deep or invisible web documents searchable without having to search each database (Chen, 2006).

In the library environment, the appropriate federated search technology to be used is Z39.50, whereas Liberty is complete web based library automation software for federated searching. Liberty offers library managers the option to include library catalogue in the federated search list and to select other databases to be included in the list. This actually makes Liberty to be customizable. On the other hand, Z39.50 is an open network standard which facilitates communication between systems that run on different hardware, with the use of different software. This makes it possible for a searcher to use the library portal's user interface to search both the local catalogue as well as any database that supports Z39.50. When a federated search is implemented in a library, it becomes a unique library service.

Mobile Library Services Within The Context of Web 3.0

The new information environment is changing rapidly and libraries are seriously under expectation to provide innovative services to meet the ever expanding needs of users. To this end, innovations in mobile library services are needed or required to meet users' needs. According to Kumar (2013), these mobile library services are an extension of digital library in the mobile communication network. Digital library and mobile communication are necessary conditions for mobile library services. Mobile technologies

such as Blackberry, iPod, Smart phones, Cell phones, Tablets, MP3 players and Personal Digital Assistants (Kumar, 2013; Saxena & Yadav, 2013) are being deployed by libraries to offer innovative and user-focused services not limited to the following:

- i. Voice service: Libraries are providing voice services with the support of mobile phones. The simplest of this service is welcoming library user into the library.
- ii. Short messaging service or text messaging: Short messages can be sent to library users through mobile phones. The messages may take the form of reminders to return library material, information on outstanding loans, library events and new resources on the library website or portal.
- iii. Multimedia messaging service (MMS): Libraries can use MMS or mobile phones to share photos, videos and radio.
- iv. Question answering service: Library users who have questions on how to use the library for a specific purpose can contact the librarians for help through mobile devices.
- v. Mobile OPAC: The Online Public Access Catalogue (OPAC) serves as a link between the user and the library collection. New mobile based OPACs are being developed to help users have access to library collection through mobile phone devices irrespective of the location of the users.
- vi. Data service: Mobile phones can now be used for web browsing and to send email to library users.

Besides, libraries can generally use mobile devices to provide practical information and bibliographical information (Ally, 2008). The practical information to be provided include address, email ids, phone numbers of the library, opening hours, rules and regulations of the library, and information regarding staff. Also, mobile library sites can provide links to quick information such as covers, news, mobile catalogue, workshops/programs and contact information. The bibliographical information is not location dependent since online or electronic bibliographical metadata can be saved or stored remotely in a server or in the cloud, and this could be accessed anytime. Users can use their mobile devices to make inquiries about books, catalogues and so on.

Prospects of Web 3.0 Applications

Libraries are growing organism which responds to evolving external environmental stimuli. The volatility of the virtual world as dictated by various phases of web technologies remains the major driver of changes in modern libraries. This is usually associated with the need to help library patrons navigate and ferret out specific information from the maze of diverse and plethora of available contents in the era of information explosion. It is in the light of the foregoing that Web 3.0 technologies are embraced in the libraries as most required development.

The semantic digital library would transcend collection of electronic books and journals to providing automatic interconnection among other fragments of hyper books from variegated sources. This enhances precision in the online retrieval system thus, leading to availability of enriched contents to users. The Web 3.0 platform makes possible the concept of virtual reality which leverages on the sophistication of smart phones and other mobile devices. This provides emergence of novel information services by merely scanning the Quick Response Codes (QR) as an interjection between the objects in the real world and the Web. Library patrons can save the author, title, and shelf number of desired library materials in their mobile phones for later retrieval and reuse purposes.

Virtual Reference Services have also become a distinguishing feature of semantic web libraries. Readable and audio contents are now transferrable to group or individual patrons to meet their informa-

tion needs as well as providing necessary digital analytics based on an artificial intelligence embedded interfaces, thereby enriching the information literacy skills of library users. The semantic web or Web 3.0 will also provide a departure from the current system of over dependence on bandwidth size for the functioning of the electronic libraries. This is because Internet is no longer restricted to computers but accessible on mobile devices. It is indeed a precursor to crystallization of libraries in the cyberspace.

The availability of WEBPACs (Web Public Access Catalogues) would be a distinguishing feature of modern libraries which is one of the components of Library 3.0. It is about remote login to the library catalogue on the cyberspace through a Uniform Resource Locator (URL). This makes the Library accessible anywhere and at a time resulting in a truly borderless library services. One of the greatest needs of the library users is the desire to collocate information from multiple online databases with divergent login details and display formats. Federated search mechanism powered by web 3.0 allows for search results to be aggregated in a single platform just like Google search. This functionality could be integrated into library management software application for effectiveness.

Web 3.0 would expand the propensity and increase adoption of cloud computing in libraries. It is becoming fashionable for academic libraries to build institutional repositories for the preservation and global visibility of their publications and research output using Dspace and E-print which are products of cloud computing. Collaborations and sharing are now mediated by online union catalogues. Libraries in Africa and other continents engage in copy cataloguing by merely login to Library of Congress and OCLC online catalogues to retrieve bibliographic records of library materials and depositing same into their server with little or no modifications.

The artificial intelligence dimension of web 3.0 makes possible the use of robots in library services. This degree of embedded system automation introduces a pick and place robot manipulator to pick books and other materials and place them in an orderly manner in pre-designated location (Lande et al, 2015). The Robots through Radio Frequency Identification (RFID) tags are capable of tracking misplaced items on the stacks, thus enhancing shelf order and speed of the retrieval system in the Library. Also through geo tagging of digital contents, library patrons with smartphones and other mobile devices having GPS (Global Positioning System) functionalities can ferret minutest information from definite locations.

Lu, Ming and Farshed (2003) summarized the prospects of web 3.0 applications within the Library context thus;

1. Library 3.0 helps librarians to provide web services to users. Web services are often regarded as one of the resources in the semantic web. Specific activities can be performed by information professionals while providing web services to users. One of such activities has to do with providing details of user interface agents that assist users on how to browse the World Wide Web. This is known as automatic web-service discovery. The next activity also involves making available to users some software agents that will interact with the web-service on specific requests or inquiries from users.

The reason for this is that multiple interactions are required between users and web services. The library users only need to tell the agent to go to amazon.com to buy a book titled 'Introductory to Economics' with not more than a certain amount of money. This is called automatic web-service invocation. Another activity by librarians involves guiding library users on how to automatically select and compose a new web-service to achieve a new objective. This process is called automatic web-service composition and interoperation. Finally, library users can be trained on how to tract and query the status of the execution of each service or request automatically. This is referred to as automatic web-service execution monitoring.

2. Information professionals are expected to utilize the potentials of the semantic web to serve the users better. The semantic web uses ontologies to describe various web resources in a structured, logical and semantic way which makes it very easy for users to have access to them. Ontologies help to bring diverse knowledge bases together and ensure that a collection of web services are equally described semantically. These knowledge bases and collections of web services are regarded as semantic web resources that provide a rewarding web experience to library users.
3. With the advent of semantic web, resources on the web are represented in ontologies. This means that ontologies can help to build semantics-based search engines. When search engines are guided by ontologies, recall and precision of search result will be increased, thus helping library users to have ease of access to various web resources. The reason is that ontologies keep content, structure and representation of an entity or knowledge agent apart. For instance, if something on “chairs” is needed or searched, two separate search results are likely to pop up. With the guidance of furniture ontology on one hand, furniture chairs will be obtained and with the guidance of person ontology on the other hand, persons who chair organizations are likely to pop up.
4. Library 3.0 promotes semantics-based retrieval of multimedia data repositories. This is important because libraries are directly having issues with the traditional content-based multimedia data retrieval due to their low features such as colour, texture and shape (Smeulders, Worring, Santini, Gupta & Jain, 2000). This problem seems to be compounded with the emergence of multi-media digital contents.

The application of web 3.0 technologies in the provision of library services provides phenomenal innovation and transformation in the entrenchment of efficiency and effectiveness in knowledge management. It gives an impressive picture and foretaste of the library of the future.

Challenges of Web 3.0 Applications

New technologies representing new opportunities often impose new challenges. These challenges differ from one web application to another. This section tries to present in a nutshell some of these challenges in developing a semantic web, managing the clouds for library services, deploying mobile communication technologies within the Web 3.0 environment and federated search systems.

Developing semantic web resources is time consuming. The reason is that the semantic web involves the creation of two formats for human and machine components. Another challenge is that the amount of information on the web today which creates concepts that are vague, contradictory and deceptive or generally imprecise and confusing. There are also other issues, such as cost, increased workload, censorship, privacy, duplications and so on, that are likely to hinder implementation of semantic web services in the library. Security and privacy are the biggest concerns with clouds management. Others issues, according to Stinivas and Chary (2017), are: the service providers can limit certain projects to certain clouds, usage tracking of multiple clouds is tedious and service providers often hide deployment details from users through virtualization.

Access authority is required for mobile communication technologies; otherwise there could be a security breach. Also, the user interfaces are varied and lack standards, which means that choice of user interface is a serious issue (Kumar, 2013).

Some databases cannot work with federated search at all, especially those ones that do not use web browser such as SciFinder scholar (Kumar et al., 2008). These authors further state that any database

that requires a login does not work with the federated search. Secondly, some databases can work with one federated search but cannot work with the other. Thirdly, the number of concurrent users for some databases is limited. This means that if someone logs into the federated search, no other person can use those databases. Fourthly, many libraries have databases on a pay-per-search basis and would not want the databases to be included in the federated search for budgetary or economic reasons. Federated search systems are basically software and as such are implemented as a service which takes quite a number of resources. Finally, in federated searching, only basic Boolean commands are used.

In their contribution, Lu et al (2003) have identified and summarized these challenges as;

1. Ontologies play critical role in organizing semantic web resources, but it is challenging to develop ontology library systems.
2. Semantic web languages have become an integral part of the semantic web as they constitute greatly what is commonly referred to as semantic web resources. The challenge is that many of these semantic web languages like RDF, RDFS do not have formal semantics. This makes it difficult but not impossible to develop semantic web. This implies that the semantics of RDF and RDFS need to be resolved by either defining formal semantics for them, or by ensuring that the problems of these languages are mainly a matter of semantics.
3. With the semantic web, anybody can say anything to anybody. Today, people can make statements that conflict. Because of this, it becomes necessary that attention is given to original statement from a known source (Proof) in order to ensure that the source is trustworthy (Trust). The challenge here is that the concept of proof and trust is yet to be formalized and a theory that integrates them into the analytics of semantic web resources is also yet to be developed.

In their summary of the challenges of emerging technology adoption in libraries, Saw and Todd (2007) identified workforce skills shortage and the librarians' stereotype and personality type as some of the barriers to the full and successful adoption of Library 3.0. They further added that libraries tend to dwell so much on the experience of the past or rather on successes made in the past, which affects the adoption of emerging technologies such as Library 3.0 model. Finally, the authors are of the opinion that institutionalization, innovation, imagination, ideation and inspiration are real issues that can fast-track the full realization of Library 3.0 model.

FUTURE RESEARCH DIRECTIONS

Effort has been made in this chapter to extensively discuss the practicability and challenges of web 3.0 technologies in the library environment. As a result, adequate light has been shed on the potentiality of the technologies towards enhanced library services. However, the chapter is not in any way claiming exhaustivity, as the subject of the research could be extend further for wider applicability. For example, taking cues from literature to examine the prospects and challenges of web 3.0 technologies relative to library services (as this chapter has done) does not have empirical basis for making definitive conclusions. Therefore, future research should focus on highlighting the benefits and possible challenges of the application of web technologies from an empirical context. Similarly, examining the phenomenon within the scope of library products and services is both inadequate and restrictive, wider applicability of the innovation from the education and business contexts should be undertaken.

CONCLUSION

Web 3.0 technologies and ontology have actually enabled Library 3.0 and have brought the tremendous change in the way information is now collected and disseminated. Library and information professionals are therefore expected to learn about web 3.0 technologies and ontology so as to become librarian 3.0 now and begin to offer engaging services to users. Library 3.0 does not only make users to browse and upload documents or user-generated contents but also to become co-creators of library services. Web technological applications are now encouraging participatory librarianship, but Library 3.0 goes beyond this to include value co-creation involving both the librarians or information professionals and library users in the information value chain. The simplest interpretation of this as espoused by Marshall and Shipman (2003) is that the semantic web plays vital role in helping human beings and computers to work together towards the organization and retrieval of information.

REFERENCES

- AlEnezi, A. K. (2012). *Library 3.0: the art of Virtual Library services*. Retrieved from http://www.academia.edu/download/44158685/Library_3.0_the_art_of_Virtual_Library_s.pdf
- Ally, M. (2008). Normadicty and information access: the mobile digital library for people on the move. In G. Needham & M. Ally (Eds.), *M-libraries: libraries on the move to provide virtual access*. Facet.
- Amrohi, R. A., Garg, A., & Chauhan, P. (2015). Application of web technology in special libraries: A challenge before library and information professionals in digital era. *International Journal of Digital Library Services*, 5(3), 43–53.
- Angadi, M., Shivarama, J., & Warriar, N. (2015). *Role of library and information professionals in Web 3.0 era*. Retrieved from: <http://ir.inflibnet.ac.in/bitstream/1944/1892/1/53.pdf>
- Balbaid, M. A. (2018). Significant benefits from libraries in Web 3.0 environment. *Journal of Fundamental and Applied Sciences*, 10(45), 327–329.
- Battacharya, A. (2016). Library 3.0 and its impact on modern library services. *International Journal of Next Generation Library and Technologies*, 2(1), 1–12.
- Belling, A., Rhodes, A., Smith, J., Thomson, S., & Thorn, B. (2011). *Exploring library 3.0 and beyond*. Retrieved from <http://www.libraries.vic.gov.au/downloads/2010-2011-shared-leadership-program-presentation-Day-1-exploring-library-3.pdf>
- Bhattacharya, A. (2016). Library 3.0 and its impact on modern library services. *International Journal of Next Generation Library and Technologies*, 2(1), 1–12.
- Burners-Lee, T. (1998). *Semantic web roadmap*. Retrieved from <http://www.w3.org/designissue/semantic.html>
- Chauhan, S. K. (2009). *Key 2 information: library 3.0*. Retrieved from <http://key2information.blogspot.com/2009/library-3.0.html>
- Chen, X. (2006). Metalib, Webfeat and Google: The strengths and weakness of federated search engines compared with Google. *Online Information Review*, 30(4), 413–427. doi:10.1108/14684520610686300

Cho, & Guistini, D. (2008). Web 3.0 and health librarians: an introduction. *Journal of the Canadian Health Libraries Association*, 29(1), 13-18.

Cohen, S., Ferreira, J., Horne, A., Kibbee, B., Mistlebauer, H., & Adam, S. (2010). Mylibrary: Personalized electronic services in the cornel university library. *D-Lib Magazine*, 6(4). Advance online publication. doi:10.1045/april2000-mistlebauer

Cox, A. M., Rutter, S., & Pinfield, S. (2018). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. *Library Hi Tech*. Advance online publication. doi:10.1108/LHT-08-2018-0105

Dent, V. F. (2007). Intelligent agent concepts in the modern library. *Library Hi Tech*, 25(1), 108–125. doi:10.1108/07378830710735894

Ekwueme, L. O., & Ilo, P. I. (2019). Marketing of library services for enhanced accessibility in national Open University of Nigeria: challenges and strategies for intervention. *Library Philosophy and Practice (e-journal)*. 2252. Retrieved from: <https://digitalcommons.unl.edu/libphilprac/2252>

Evans, W. (2009). *Building library 3.0*. Chandos. doi:10.1533/9781780631868

Eysenbach, G. (2008). Medicine 2.0 social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10(3), e22. doi:10.2196/jmir.1030 PMID:18725354

Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E. K., & Stephens, S. (2009). *The semantic web in action*. Retrieved from <http://www.scientificamerican.com/article.efm?id=semantic-web-in-action&page=1>

Giustini, D. (2007). Web 3.0 and medicine. *British Medical Journal*, 335(7633), 1273–1274. doi:10.1136/bmj.39428.494236.BE PMID:18156223

Govanakoppa, R. A., & Kumara, B. (2014). *Web 3.0 tools: applications and impact on library services, in the book of world in digital age: digitization and preservation*. INMANTIC Publishing house.

Herdler, J. (2008). Web 3.0: chicken farms on the semantic web. *IEEE Computer*, 41(1), 106-108.

Hielmcrone, H., Maiello, R., Bainton, T., & Bonnet, V. (2012). E-publishing and the challenges for libraries: An information paper. *European Bureau of Library, Information and Documentation Associations (EBLID)*. Retrieved from: www.eblida.org/Experts%20Groups%20papers/EGIL-papers/02_EBLIDA_E-publishing%20and%20the%20challenge%20for%20libraries_Background_information_Public_version.pdf

Izuagbe, R. (2017). Pricing implication of publishers' licensing models on sustainable e-resources integration in academic libraries: A review. *International Journal of Information Science and Management*, 15(2), 79–88.

JastramI. (2008). *Library 3.0*. Retrieved from: <http://pegasulibrarian.com/2008/08/library3.0html>

Kadyam, S., & Singroha, R. (2014). Web 3.0 in library services: An utilitarian effect. *Journal of International Management*, 1(2), 159–166.

Kang, J. Y., & Yong, H. S. (2007). Web 2.0 concepts and perspective. *Communications of the Korean Institute of Information Scientists and Engineers*, 25(10), 5–11.

Kaur, D., & Sharma, P. (2017). Effectiveness of web-based social sensing in health information dissemination—A review. *Telematics and Informatics*, 34(1), 194–219. doi:10.1016/j.tele.2016.04.012

Kumar, R. S. (2013). Mobile phones: A new dimension in library services. *International Journal of Library Science*, 7(1), 39–46.

Kumar, S., Sanaman, G., & Rai, N. (2008). Federated search: new option for libraries in the digital era. In *International CALIBER* (pp. 267 – 285). Academic Press.

Kwanya, T., Stilwell, C., & Underwood, P. G. (2012). Intelligent libraries and aopmediators: Distinguishing between library 3.0 and library 2.0. *Journal of Librarianship and Information Science*, 45(3), 187–197. doi:10.1177/0961000611435256

Libner, K. (2003). *Working the network: a future for the academic library*. Retrieved from <http://alpha.fdu.edu/~marcum/libner.doc>

Lipow, A. G. (2005). *The virtual reference librarians' handbook*. Neal-Schuman Publishers.

Lu, S., Ming, D., & Farshed, F. (2003). The Semantic Web: opportunities and challenges for next-generation web applications. *Information Research*, 7(4). Retrieved from <http://ifnformationR.net/ir/7-4/paper134.html>

Lucier, T. (2009). *What is web 3.0 definition*. Retrieved from <http://www.tourismkeys.ca/blog/2009/01/what-is-web-30-definition>

Marshall, C. C., & Shipman, F. M. (2003). *Which semantic web?* Retrieved from <http://www.csdi.tamu.edu/~marchal/ht03-sw-4.pdf>

MetzC. (2007). *Web 3.0*. Retrieved from <http://www.pemag.com/article2/0,2817,2102852,00.asp>

Morris, R. D. (2011). Web 3.0: Implications for online services. *TechTrends*, 55(1), 42–46. doi:10.1007/11528-011-0469-9

Nesta, F., & Mi, J. (2011). Library 2.0 or Library III: Returning to leadership. *Library Management*, 3(1), 85–97. doi:10.1108/01435121111102601

Nkiko, C. (2014). *Managing ICT-driven libraries: critical leadership issues*. Leadership and Innovation: Cutting Edge practices for the 21st Century Librarian Conference, Covenant University, Nigeria.

Noh, Y. (2015). Imagining library 4.0: Creating a model for future libraries. *Journal of Academic Librarianship*, 41(6), 786–797. doi:10.1016/j.acalib.2015.08.020

O'Connor, D. (2010). Apomediation and ancilliary care: Researches' responsibilities in health-related online communities. *International Journal of Internet Research Ethics*, 3(12). http://ijire.net/issue-3.1/7_O'connor.pdf

Pinto, M., & Manso, R. (2012). Virtual references services: Defining the criteria and indicators to evaluate them. *The Electronic Library*, 30(1), 51–69. doi:10.1108/02640471211204060

Rao, N. K., & Babu, K. H. (2001). Role of librarian in internet and World Wide Web environment. *Informing Science*, 4(1), 26–34.

Ruthman, D. (2008). *Disliking 'web 2.0' and hating 'web 3.0'*. Retrieved from <http://davidrothman.net/2008/01/08/disliking-web2.0-andhating-web3.0/>

Sapp, M. N., & Van Epps, A. S. (2006). *Equal access: what does the digital revolution mean for library websites*. Retrieved from <http://does.lib.purdue.edu/lib-research/30/>

Saw, G., & Todd, H. (2007). *Library 3.0: where art our skills?* Paper presented at the world library and information congress: 73rd general conference and council, Durban, South Africa. Retrieved from <http://archive.ifla.org/iv/ifla73/papers/151-Saw-Todd-en.pdf>

Saxena, A., & Yadav, R. D. (2013). Impact of mobile technology on libraries: A descriptive study. *International Journal of Digital Library Services*, 3(4), 1–13.

Schultz, W. (2006). To a temporary place in time: on the way to the library experience of the future. *OCLC Newsletter*. Retrieved from <http://www.oclc.org/nextspace/002/6.htm>

Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1349–1380. doi:10.1109/34.895972

Stinivas, K., & Chary, K. D. (2017). Key issues and challenges in cloud computing for library services. *International Journal of Engineering Research and Application*, 7(12), 80–82.

Strickland, J. (2010). *How web 3.0 will work*. Retrieved from <http://www.hardyinst.com/newsletter/dec2010/web3.0.pdf>

Wahlster, W., & Dengel, A. (2006). Web 3.0: Convergence of web 2.0 and the semantic web. *Technology Radar*, 2, 1–23.

KEY TERMS AND DEFINITIONS

Cloud Computing: A model for information network, storage, and access, facilitated by a remote server hosted over the internet instead of a local server or computer.

Library 3.0: A computer-mediated library that promotes access to a wider array of information resources and services beyond its physical location.

Library Services: Information-based services offered by the library to a defined group of patrons for personal self-development, educational, business, political and other purposes with the aid of man and or technology infrastructures.

Mobile Technologies: Hand-held internet-compliant devices that foster wireless communication and computing applications using the Global System for Mobile (GSM) network which include android phones, iPhones, iPads, and tablets.

Semantic Web: A simplified world wide web that fosters human and computer interaction for effective management of data/information.

Virtual Reference Services: Electronic means of communicating and conducting referrals and other reference-based services.

Web Technologies: Mode of communication between or among interconnected computers using hypertext markup languages and other multimedia applications.

Massive Digital Libraries (MDLs) and the Impact of Mass–Digitized Book Collections

Andrew Philip Weiss

 <https://orcid.org/0000-0002-8900-2779>

California State University, Northridge, USA

INTRODUCTION

To better analyze the growth of digital libraries, Weiss and James have proposed adopting the term Massive Digital Libraries (MDLs), a concept based on the increased size, scope and scalability of mass-digitized book collections. MDLs rival physical libraries' print holdings in size, breadth, and depth, often approaching a scale previously found only among library consortia or national libraries. (Weiss and James, 2013a, 2013b, 2014, 2015; Weiss, 2016) The concept further intersects the digital library with the wider development of 'big data,' which is driven to an ever-larger scale by the '5Vs' of Volume, Variety, Velocity, Variability, and Veracity. (Weiss, 2018)

The root of the concept begins in late 2004 when Google publicly announced its intention to digitize the world's books—including works still under copyright protection—and to place them *all* (roughly 130 million) online. Jean-Noel Jeanneney, head of Bibliothèque nationale de France at the time, interpreted Google's planned project as a wake-up call for European countries. Failure to catch up to the American company, he argued, would result in significant problems for non-American organizations, especially if they were unable to check the company's outsized influence. (Jeanneney, 2005)

Fifteen years on, it is hard to imagine that Google's desire to create an online digital should have come as such a shock. Yet Google spurred significant hand-wringing and soul-searching among institutions traditionally charged with producing or preserving cultural artifacts. (Jeanneney; Venkatraman, 2009) In retrospect, the controversy seems quaint in comparison to the current crop of issues – especially the current “disruptions” of established economic models by Uber/Lyft, Facebook, Twitter, Spotify, and the like; the looming unknowns circling around artificial intelligence; and the encroachments on civil rights via electronic digital surveillance and other intrusions of privacy.

A number of similar mass-digitization projects developed and matured since Google's announcement, including the *HathiTrust*, *Internet Archive*, *Digital Public Library of America (DPLA)*, *California Digital Library*, *Texas Digital Library*, *Gallica*, and *Europeana*. These projects each transcend their roots as localized digital libraries and simultaneously adapted to and altered the digital landscape. These various MDLs have allowed for and contributed to the ascendancy of our current mass-digitization online culture.

This chapter describes the characteristics of Massive Digital Libraries (MDLs) and outlines their impact upon current information science issues, especially with regard to digital collection metadata, copyright, the diversity of source collections, and user privacy in an age of so-called “surveillance capitalism.” (Zuboff, 2015) Traditionally, libraries have been created to serve particular communities *defined* as well as *bound* by geography, intellectual disciplines, or specific end users. MDLs in their current trajectories, however, promise to transcend such limits in ways that are simultaneously *constructive* and *destructive*.

BACKGROUND: DEFINING MASSIVE DIGITAL LIBRARIES

Defining Criteria

Massive Digital Libraries (MDLs) describes a specific class of digital libraries that correspond to the size of a traditional, large brick-and-mortar library. Although other disciplines have discussed digital libraries and archives in terms of computer science, such as in the Very Large Digital Library (VLDL) movement, none have framed the discussion in terms of the principles of library science or the services and content access provided by an actual, working library. (VLDL, 2011)

The following list of characteristics has been proposed to help define MDLs:

1. Collection size: surpasses 500,000 texts; prime MDLs comprise tens of millions of texts;
2. Acquisitions, collection development & copyright: numerous partnering members contribute content regardless of author or copyright holder permissions and regardless of end-user needs;
3. Content type: mass-digitized print books; the resulting searchable digital corpus of texts becomes as important as the individual titles;
4. Collection diversity: diversity is dependent upon self-chosen partner members, which can reflect distortions or biases inherent to the source collections;
5. Content access: varying degrees of open access exist within MDLs; content is searchable through single, uniform interfaces (search engines & portals) representing all the collections as members of a single entity regardless of source;
6. Metadata: Metadata is gathered and aggregated from multiple sources, with a reliance on new digital description schema;
7. Content / digital preservation: consortium members provide long-term digital preservation strategies at local levels as well as “in the cloud”.

These criteria and their attendant issues, though not necessarily *unique* to digital libraries, require different approaches when dealt with as a Massive Digital Library. The issues involved with aggregating millions of previously published print materials into one uniform, yet decentralized, conceptual and online digital space become more complex as the size increases. It is important to differentiate MDLs from their smaller counterparts as they are more difficult to police and analyze, especially with metadata uniformity, copyright compliance, rights ownership, and user privacy. The larger an institution or system is, the more unwieldy and slow-to-change it may become. Furthermore, the lack of transparency among IT companies – especially Google – requires more robust oversight. As a result, it is important to remain cognizant of how MDLs approach the common characteristics of books in ways that stretch the boundaries of the original print medium. Print books remain static during their lifetimes, changing only when new editions are created. E-books promise the flexibility to change into ongoing “works in progress,” without fixed texts. Unlike print books, their access can easily be tracked, eroding the user privacy that remains a central tenet of libraries. Identifying MDLs as unique entities would allow scholars and researchers to both utilize and safeguard these newly-created and widely-distributed digital collections.

Representative MDLs

The following is a look at several representative MDLs and their defining characteristics. This section will also help readers get a sense of how MDLs stack up against each other. Criteria described for each

MDL include their starting dates, *modus operandi*, estimated sizes of the collections, subjects collected, the number of partner institutions, and the languages covered in their collections.

I. American-Based, Anglo-Centric MDLs

A. Google Books (25 million texts)

The *Google Books* project started in 2004. Envisioned by Google co-founder Sergei Brin as a replacement for a library's 'card catalog,' *Google Books* currently holds approximately 25 million volumes within its collections, though exact figures have been difficult to verify. Google has not provided updated statistics on its collection's size in several years, and has mostly waned in its efforts to maintain the collection. As a result, 25 million may be an inaccurate figure.

Google Books is characterized in part by its extremely robust and flexible search engine. When searching in this MDL, a large amount of set results will often appear. Mistakes in the search are often auto-corrected. Their stated goal is the digitization of all the books in the world, a number that reaches about 129,864,880. (Jackson, 2010) They work with authors directly through their Google Books Partner Program and with public organizations through their Google Books Library Project, which includes about 22 public and academic institutions and consortia from several countries, including Japan, Germany, England, and Spain. (Sato, 2010) Overall collection diversity is represented with a large number of languages represented and accessible, though the bulk of the source material in the digitized corpus is in English.

The *Google Books* corpus is open to anyone able to use a computer that is connected to the Internet. However, due to copyright restrictions and the threat of litigation, any materials falling outside the public domain have been closed to full-text viewing. Google provides users with three levels of accessibility to the texts beyond the basic metadata record: 1) Snippet View, which is a few sentences of the book, 2) Partial view, which is often several to dozens of pages, and 3) Full view, which is the complete text. Additionally, Google has provided access to the data sets from the digitized corpus for researchers to utilize. One of the most promising tools developed for this text mining is the Google Ngram viewer.

B. HathiTrust (17 million Texts)

HathiTrust started in 2008. It currently holds 17 million volumes (representing 8.2 million unique book titles) within its collections; 6.5 million volumes (38% of the total) are in the public domain. The *HathiTrust's* main goal is to meet the needs of the members of its partnering institutions. These partners include nearly 72 world-class academic research and national libraries. Notable partners include the Library of Congress and the California Digital Library. Collection development is focused on gathering digitized books and journals from partner institutions and in-house digital initiatives and those books digitized in part by *Google Books* and the *Internet Archive*.

As one might expect, English is the main language of the archive. Roughly 4.6 million of their books, representing 51% of the unique holdings, are in English. The second largest represented group is German (8.6%), followed by French (6.9%), Spanish (6.5%), Russian (3.3%), Chinese (3%), Japanese (2.7%), and Italian (2.5%). Given the large Spanish-speaking population in the United States, the small and disproportionate amount of Spanish-language materials is surprising. However, given that the books primarily come from research universities' collections, German and French (second and third respectively) have

historically represented the cutting-edge of scholarship in more disciplines than Spanish. (HathiTrust, 2019) This disparity will be discussed in greater detail below.-

HathiTrust is characterized by a robust search engine and extremely well-organized metadata and records management. Where *Google Books* has issues with the retrieval of specific dates and other types of metadata that significantly hamper its usefulness, *HathiTrust* performs information retrieval and document display with superior results. The system allows, however, only two types of views to the non-member user: 1) a limited view, which provides only search results to retrieve word counts from queries (i.e. full-text search), and 2) full-view, which provides the full text – but this is reserved primarily for materials in the public domain or those with specific author-provided permissions for open access.

Access is open to anyone able to use a computer connected to the Internet, but partnering institutions and their member users benefit from extra services such as content storage, data preservation and copy-right clearance/usage consultation. Although the current size of the collection is smaller in comparison to *Google Books*, the ability for one to find information more easily should make the *HathiTrust* the go-to MDL for *all* libraries and their patrons. Indeed, the collection size is slowly but steadily catching up to *Google Books*.

C. Internet Archive (15 Million Texts)

The Internet Archive started in 1996 as web archiving service. It currently holds 15 million volumes of texts. It is a unique addition to the Massive Digital Library category. On one hand their main focus appears to be on archiving the Internet via its *Wayback Machine*. One of their stated goals, in fact, is to create archival artifacts of the web itself, and they have archived an astounding 330 billion webpages. Unlike *Google Books* or *HathiTrust*, which focus heavily on book collection sizes, *Internet Archive's* primary concern has been less focused on the number of titles acquired. Instead, the *Internet Archive* gathers and approves *only those materials in the public domain or that have explicit author permissions*. As a result, their print-derived digital collections are slightly smaller but completely accessible.

Their advantage is also in providing links to several mass-digitization projects and allowing these projects to exist under their own names as well as within the Internet Archive web pages. Projects such as the *Biodiversity Heritage Library*, *Project Gutenberg*, the *Universal Library* and *Children's Library* are worthy open access book digitization initiatives aimed at providing people with free access to resources, all falling under the *Internet Archive* umbrella.

Their difference is also in one of scope, with their emphasis on audio recordings, video, electronic video gaming, and other unique multi-media collections. They claim to have 4.5 million audio recordings (including 180,000 live concerts), 4 million videos (including 1.6 million Television News programs), 3 million images, and 200,000 software programs in their collections. Additionally, the *Internet Archive* provides access to about 9,500 audio recordings of books, including literature and poetry. (Internet Archive, 2019)

D. Other Anglo-Centric MDLs of Note

In recent years several other U.S.-based MDLs of significance have begun to appear. The most well-known of these, The *Digital Public Library of America* (DPLA) is also one of the newest. Due to aggressive partnership and digital hub development, its growth since October 2010 has been exceedingly quick. It currently holds nearly 35 million digital items, a huge increase from the 2.5 million at its initial launch, and nearly 16.2 million texts. The collections specialize in a range of materials from books to

digital archives. The scope ranges from materials for students and teachers at the K-12 level, and for universities up to the graduate-student level. The main partnering institutions include public libraries, university libraries, historical societies, museums, and archives. It is now a major up-and-coming MDL and has clearly learned from the negative experiences and outcomes of its predecessors. (DPLA, 2019)

Additionally, the *California Digital Library* and *Texas Digital Library* continue to grow. Although they serve more localized, state-based interests, their open online accessibility ensures widespread impact. Many of CDL's digital book collections appear as subsets within Google Books, the HathiTrust, DPLA, and Internet Archive. (CDL, 2019) The Texas Digital Library's collections focus on large-scale institutional repository development and issues inherent to scholarly communication, especially ETDs, data management planning and archiving, digital preservation and digital objects. (TDL, 2019)

II: NON-ANGLO-CENTRIC MDLS OF NOTE

A. Europeana (58.7 million Items / 12.1 Million Texts)

The *Europeana* digital library aggregates multi-media content from nearly 240 institutions across Europe, including video, image, text and digital artifacts. It is a multi-lingual, multi-cultural endeavor. There appears not to be one dominant language in the project. With a huge collection of over 12.1 million texts the digital library should be considered a leading MDL in the field. (Europeana, 2019) Their emphasis on interoperability and a strong metadata schema allows users to refine searches and hone in on more relevant search results easily. The main focus for the library is to ensure that digital materials of numerous cultures remain freely accessible to all. The project advocates the strongest pro-Public Domain stance of nearly all the MDLs. (Fallon, n.d.) It is a vital and important MDL as a result.

B. Gallica (5 Million Items / 695 Thousand Texts)

Despite its smaller book collection size (less than 1 million), *Gallica* is noteworthy for several reasons. One, it was the first digital library developed in direct response to the Google Books project. (Gallica, 2019). Two, it is primarily and uniquely a France-centric, Franco-phone Massive Digital Library. Its partners include many universities and museums in France but also the National Library of Brazil and the Library of Congress. There are books offered in English and other languages, but it is a small amount compared to the size of the French language collection. *Gallica* is also, in many ways, an MDL counterpoint to *Google Books*, with a very different philosophy and approach to the digital library.

III. MAJOR ISSUES AND CONTROVERSIES

A. MDLs, Mass Aggregation & The Accuracy and Veracity of Results

Google Books' policy for developing their Massive Digital Library appears to have been very simple: *digitize everything*. Such ambition to collect all of the world's knowledge is not unheard of in the history of libraries, yet these are time-consuming and costly efforts requiring extreme amounts of funding and personnel to do them adequate justice. It is as ambitious as it is unrealistic. It is possible complete digitization may never happen as new books are still being published in print form yearly. Furthermore,

not all books are available with digital counterparts or can even be scanned accurately and adequately. Some works made with high-quality materials, odd-sized or hand-crafted paper, or three-dimensional forms, are preferable to many readers as solid objects and not as screen-bound 2D digital copies. Some countries such as Japan have traditional binding procedures that are unique and not easily transferrable to digital formats.

The estimation of 129 million books may also be flawed. Relying on ISBN numbers is not accurate as many countries did not adopt ISBNs until the 1960s or 1970s. The differentiation between serials, multi-volume works, multiple editions and the concept of “books” is unclear as well. For example, does Google consider the numerous volumes of the Harvard classics as one book series, with one record as in a library catalog or the Hathitrust, or are they conceptualized as separate books? Does the second edition of volume 13 of the Harvard Classics count in the tally if the first edition is available? Furthermore, is it vitally important to have a scanned copy of every edition of a classic text? A “digitize everything” approach does not suitably answer these questions and an assumption of universality becomes a disservice to all.

Endeavoring to digitize all the books is not so much a collection policy as it is an *avoidance* of one on a massive scale. The results are akin to a typical Google keyword search: a mixture of high and low culture with only relevance ranking, which can be manipulated by both black and white hat Search Engine Optimization (SEO), to determine whether it is a satisfactory result. One concrete example of this issue has been the inability for the Google Books search engine to accurately retrieve and coherently display single works of multiple-volume book series. A search for <Harvard Classics> results in a jumbled non-sequential retrieval of titles. Unless one had known before-hand that the *Harvard Classics* series contains 51 volumes, one might never get an accurate accounting of the texts from Google’s set result.

At the same time, digitizing all content will also include books of questionable natures, including the literatures of hate, pornography, and other provocative or even dangerous materials. Yet, despite their intention to digitize everything, it is unclear due to Google’s secrecy whether they have digitized materials that are hateful or offensive to various groups.

One must be wary of this approach as well, because not only will books be missed, especially those not held by Google’s Partners, but also a lot of unnecessary, spurious, or poorly written materials will be added to the collection. This in turn dilutes the rationale for collecting something.

Additionally, the accuracy of the metadata needed to track and provide information retrieval in meaningful human-readable ways is also at risk. James and Weiss (2012) found a high rate (36%) of errors in their study of Google Books’ metadata. Other scholars have been pointing out the “train wreck” that is Google Books’ metadata for quite some time. (Duguid, 2007; Nunberg, 2009) Weiss and James also examined rates of error among metadata and in scanning results. James (2010), Weiss and James (2013a, 2013b, 2014, 2015), and Weiss (2016) have found significant evidence of errors in Google Books.

All of these issues – from responsible collection development and content aggregation to the accurate retrieval of information – are intertwined. Including reputable and appropriate sources, accurate information, and verifiable content instills trust in users, something traditional libraries have done very well. Google Books, on the other hand, has struggled with this. Without trust, users will eventually abandon the service.

B. MDLs & Copyright

MDLs have greatly impacted copyright law. MDLs can be seen as logical extensions of the Internet’s early promise of the information superhighway and the democratization of knowledge. Yet there has been intense backlash from all sides against MDLs. Indeed, the two largest of the Massive Digital Libraries,

HathiTrust and *Google Books*, were saddled with high-profile lawsuits for years, both initiated by the Author's Guild, from the mid-2000s until 2015.

By November, 2013, Google's long legal fight eased somewhat with the ruling in Google's favor by Judge Denny Chin. (Lee, 2013; Chant 2013) By 2014 appeals were also found in favor of the HathiTrust over the Authors Guild. (Chant 2014) By the end of 2015 Google's case had been resolved with the United States Second Circuit ruling in the fall of that year in favor of the tech giant, and the U.S. Supreme Court declining to examine the case further. (Hurley, 2016)

Looking at it more objectively, though, Google's legal problems stemmed from the proprietary nature of their endeavor. If they had been willing to abandon some of the profiteering and posturing, or to alter the purpose of their program, the project might have been seen in a better light by the Authors Guild. Despite Google's controversial approach, however, the U.S. Supreme Court considered the project to fall within the guidelines of fair use, especially regarding the ultimately transformative use of the source material. (Hurley)

HathiTrust, in contrast, weathered the copyright lawsuit storm quite well and remains poised to develop as the leader of MDLs in terms of copyright compliance and providing accessibility to the disabled and access to public domain works. The result of their lawsuit showed that the fair use defense remains a strong one as long as it is coupled with such factors as restricting access to copyrighted work and remaining non-profit in nature. Their digitized corpus was also ruled fair use because of the importance the *HathiTrust* placed upon providing works with greater accessibility for the disabled. The fair use defense also made it clear that the *scope* of a project, which the Author's Guild had criticized, is irrelevant. Fair use holds in principle regardless of the number of books they digitized.

Among other MDLs, *Europeana's* efforts in advocating for a strong and robust public domain make it a worthy cause to follow. (Fallon, n.d.) It is hoped that they continue to fight for strong and clear open access of cultural content disambiguated from the profit motives of the corporations that tend to own the lion's share of copyrighted material.

C. The Promises and Pitfalls of Diversity

Google Books' partners have included institutions in Spain, Germany, France, and one university in Japan. Given the huge number of books, languages and cultures in the world, one must consider deeply whether such limited partnerships are feasible and scalable. In order to have complete universal coverage, one would need to add partners from everywhere, including institutions devoted to sub-communities and sub-cultures within larger countries. But despite such ambitions to create a digital corpus of everything printed, it is likely not a realistic outcome.

Furthermore, the act of scanning itself causes distinct issues. As Thylstrup writes, once a decision is made to digitize content, scanning becomes a political act rife with "distinct subpolitical processes." (2014) In addition, researchers have found that the contents of the Google Book corpus do not represent an unbiased sampling. Pechenick, et al. argue that "the evolution of the corpus throughout the 1900s is increasingly dominated by scientific publications rather than popular works. We have shown that even the first data set specifically labeled as fiction appears to be saturated with medical literature." (2015) Aside from the problem of misrepresenting common collections, or transforming the boundaries of traditional print forms, the selection of texts appears to be skewed in favor of scientific titles. Weiss and James find missing public domain books in the Google corpus and a low representation of Spanish and Japanese language texts. (2013b, 2015) The diversity of content in the *Google Books* project as well as any of the other MDLs, therefore, remains at the mercy of the partnerships they have created.

For the *HathiTrust*, similar issues exist. The number of partners is large, but still not comprehensive. According to their online statistics, the percentage of English language books remains at about 50%. Of the top ten languages, 60% are Indo-European, with one, Latin, being a dead language. Two are East-Asian Languages (Chinese / Japanese) and one is a Slavic language (Russian). (HathiTrust, 2016) By world population and the number of cultural artifact created by various cultures – especially those that are older than the United States by millennia – the collections are disproportionately skewed toward the English language. Again, like the *Google Books* project -- and in part because many of the *HathiTrust* partners as well as digitized books are shared with *Google Books*--the issue of the source materials calls into question the ability of MDLs to create collections reflective of an increasingly diverse population.

For Public Domain works, this proportion skews even further to English - about 60 percent of titles in the public domain are English with 40 percent left for any other language. Generally, this reflects the needs of the faculty and students at the academic institutions currently partnering with the *HathiTrust*. It is quite telling that the second and third largest collections of works are German and French, while Spanish, the second most-spoken language in the United States, trails at a distant 4th. Only 6.5% of the HathiTrust's online collection is in Spanish, yet 36 million people (roughly 12% of the population) in the United States speak Spanish as their *first* language. In contrast, only about 1.3 million (or .41%) speak French and only about 1.1 million (about .35%) speak German, yet the HathiTrust collections include 7% and 9.2% of these languages respectively. (US Census Bureau, 2011)

It is clear, then, that subject diversity could be better addressed by even the MDL exemplars. It is hoped that as these digital libraries develop over time, the gaps found in their coverage will adapt to the new, at-large audiences that exist beyond the physical boundaries of the locations they represent.

D. Privacy and The MDL

Finally, one of the under-examined issues regarding Massive Digital Libraries is the problem of ensuring and safeguarding user privacy. User privacy is one of the pillars of library policy, especially regarding the subject matter that users search for. Safeguards exist locally, regionally, and nationally in terms of information, especially with medical and student information (i.e. HIPAA, FERPA). Libraries follow this as well with local user policies that are backed up or enforced by guidelines from state and national library and archive organizations. The American Library Association, for example, provides clear policy for libraries nationwide, regardless of their constituencies. (ALA, 2014)

While HathiTrust, DPLA, CDL, and TDL closely follow privacy policies supported by the ALA, the same cannot be said for Google. Indeed, one of the main issues with the tech company has been its well-publicized disregard for how user data is handled. After it is collected, it is often shared or sold to third-parties for advertising revenue. This type of “surveillance capitalism” as defined by Shoshanna Zuboff, greatly compromises the benefits the Google Books corpus provides for researchers and users. It is unclear how or to what extent Google has shared or sold the data about its users of the Google Books database, but doing so flies in the face of a century of modern library practice. Big data surveillance meets the digital library.

IV. SOLUTIONS AND RECOMMENDATIONS

Regarding the problems of faulty scanning, digital representation and verisimilitude, it is likely that improved algorithms and AI for image correction may improve the problems related to book digitization.

AI may also be used to help with improving the rates of error in metadata associated with the books. In the case of Google Books, they would be well-advised to drop the BISAC-derived subject headings, which are used by booksellers and not libraries, and take efforts to use the metadata and controlled vocabularies developed by libraries over decades of intense and precise practice. This might help with improving the generally dismal experience of Google Books.

Google's cavalier approach to copyright should be taken as a cautionary tale. HathiTrust's approach to copyright, rigorous checks and adherence to copyright permissions, on the other hand, *should* be emulated by all MDLs. Their innovative copyright clearance project uses hundreds of volunteers to manually research and double-check the copyright status of many orphan works. While it is a widely-known rule of thumb that a book published prior to 1924 is in the public domain, it is less well-known that many works published after 1924 may actually be in the public domain as a result of various factors, including non-compliance with rules and failure to re-register a work after its initial copyright period. The HathiTrust's copyright verification project carries out the due diligence necessary to make sure scanned works can be released to the general public. Notably, U.S. copyright law in January 2019 finally allowed new works into the public domain after a twenty year freeze. MDLs will be well-advised to open these previously closed works. Each year after this, works will be released.

Finally, libraries and MDLs need to be advocates for and practitioners of stronger online privacy laws. Current user-tracking practices erode online privacy for the sake of profits derived from the basic tenets of "surveillance capitalism," which "exile persons from their own behavior while producing new markets of behavioral prediction and modification." (Zuboff, 2015) The dampening impact of surveillance on information-seeking behavior cannot be underestimated. MDL governance structures and governing bodies will need to be vigilant advocates for better privacy safeguards, especially as Google and Google Books continue track users' every move. The European Union's *General Data Protection Regulation* (GDPR) is a good first step, but more widespread policies across all nations must be adopted.

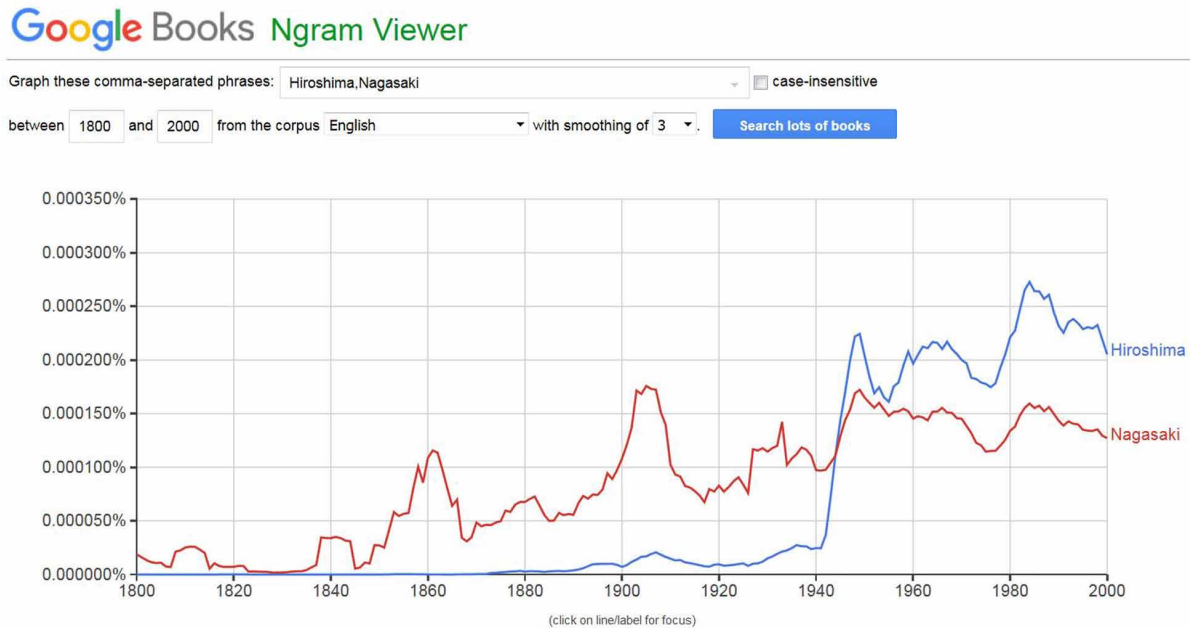
V. FUTURE RESEARCH DIRECTIONS: THE IMPACT OF DIGITIZED BOOKS

A. Humanities and the Ngram Viewer

One important impact that MDLs can have on current scholarship is the mining of the immense data sets generated from the mass digitization process. The Ngram viewer, designed by John Orwant and Will Brockman at Google, provides data visualizations for the frequency of queried terms. This is a helpful tool for not only historians, writers, artists, and social scientists, but also libraries. (Google N-gram, 2016) Michel, et al. (2011) have shown the impact that research in this area, which they have called "Culturomics," can have on scholarship and textual analysis. Additionally, one can track historical events in the news or the ideas and concepts that accompany them. It is truly a useful tool for those interested in the history of ideas and how events shape them.

In the example shown in the figure above, a search involving the terms <Hiroshima> (in blue) and <Nagasaki> (in red) was conducted by the author. The search results show the frequency with which the words appeared in the English language from the period of time from 1800 to 2000. For the first 140 years, Nagasaki was the more mentioned of the two cities, likely a result of it being the only port open to foreign trade during the closed-country *sakoku* period in Japan. During the 1850s-1860s, Nagasaki also began to surge, likely a result of the United States' aggressive policies to "open" up the country to international trade. Again a surge occurred between 1900 and 1910, reflecting its status as a major hub for coal production and the Japanese Navy's winter port for the Russian Asiatic fleet.

Figure 1. Ngram viewer showing frequency of the names Hiroshima (blue) and Nagasaki (red) from 1800 to 2000. Note the significant change starting in the early 1940s and spiking post-WWII.



Hiroshima during this same period of time is not mentioned very much in comparison, trailing at a very distant second. However, during the years after the bombing of Pearl Harbor, Hiroshima becomes ever more frequently mentioned. Only by 1944, a year prior to the atomic bombings of both cities, do Hiroshima and Nagasaki reach parity in terms of frequency. After 1945 until the present day as a result of being the first city to suffer an atomic bomb attack, Hiroshima has been mentioned the most.

It is a useful tool for students to get a better sense of the context of historical events and an understanding of how notoriety waxes and wanes over time. It becomes clear then that MDLs can offer more than just digital versions of books and get at assisting not only scholars but also interested lay-persons, librarians, and students. The ability to search for concepts across a large sample of materials will revolutionize the humanities, the history of science, and even science itself, as people begin to see the patterns in knowledge development and information retrieval.

CONCLUSION

The Massive Digital Libraries covered in this chapter show striking individuality in their approaches to mass-digitization and the mass-distribution of digitized content. Yet many of their defining characteristics – from size to content aggregation, to development of consortiums and to content diversity and general accessibility – nonetheless remain similar. Their central missions to share the content of print books that may be locked away in print on physical book shelves are also admirable. Varying degrees of scope, subject matter and audience involvement shape these massive collections in individual ways, making each MDL unique.

Overall when looking at the main examples of MDLs currently in existence, one gets the sense that ambitions for a “unified digital front” for all types of materials is the impetus driving their development. This ambition impacts brick-and-mortar libraries in that they may have to scramble to accommodate as well as take advantage of these growing initiatives.

The issues raised by MDLs speak directly to the issues at the heart of mass-digitization culture. Veracity, accuracy and diversity get at the essential quality of any information society: trustworthiness. If one is able to trust the quality and comprehensiveness of these mass-aggregated collections, then MDLs will provide an essential service to the pluralistic and democratic societies from which they have sprung. Additionally, the issues of copyright are providing a lesson to all on fair use and the development of new open access/open source cultures dependent upon the alteration and recombination of previously created works. Privacy needs to be guarded for the user, however; otherwise, users will rightly see MDLs as intrusive tools that have the ability to oppress those searching for unpopular or controversial ideas. The new tools and applications of MDLs, such as the Ngram Viewer, despite some of the issues of representation existing at the heart of the Google Books corpus, promise to augment the value of individual books beyond their physical limitations. MDLs will ultimately provide users with enhanced approaches to existing disciplines as well as the promise of new transformations.

REFERENCES

- American Library Association (ALA). (2014). *Privacy: Advocacy, Legislation and Issues*. <http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy>
- Archive, I. (2019). *About the Internet Archive*. <https://archive.org/about/>
- California Digital Library (CDL). (2019). *Home page*. <https://www.cdlib.org/>
- Chant, I. (2013). Lawsuit against Google Books dismissed. *Library Journal*, 138(20), 14.
- Chant, I. (2014). Academic: Appeals court upholds HathiTrust verdict. *Library Journal*, 590(12), 15.
- Digital Public Library of America (DPLA). (2019). *Digital Public Library of America Home*. <https://dp.la/>
- Duguid, P. (2007). Inheritance and loss? A brief survey of Google Books. *First Monday*, 12. doi:10.5210/fm.v12i8.1972
- Europeana. (2019). *Europeana Collections*. <https://www.europeana.eu/portal/en>
- Fallon, J. (n.d.). Europeana and the Public Domain. *Europeana Professional*. <https://pro.europeana.eu/public-domain-content>
- Gallica. (2019). A propos. *Bibliothèque Nationale du France. (BNF)*. <https://gallica.bnf.fr/html/und/a-propos>
- Google. (2013). Google Ngram Viewer. *Google Books*. <https://books.google.com/ngrams>
- HathiTrust. (2019). *HathiTrust Languages*. https://www.hathitrust.org/visualizations_languages
- Hurley, L. (2016). *Reuters. Supreme Court rejects challenge to Google book-scanning project*. <https://www.reuters.com/article/us-usa-court-google-idUSKCN0XF1O3>

Jackson, J. (2010, August 6). Google: 129 Million Different Books Have Been Published. *PC World*. https://www.pcworld.com/article/202803/google_129_million_different_books_have_been_published.html

James, R. (2010). An assessment of the legibility of Google Books. *Journal of Access Services*, 7(4), 223–228. doi:10.1080/15367967.2010.503486

James, R., & Weiss, A. (2012). An assessment of Google Books' Metadata. *Journal of Library Metadata*, 12(1), 15–22. doi:10.1080/19386389.2012.652566

Jeanneney, J. (2007). *Google and the myth of universal knowledge: a view from Europe*. University of Chicago Press.

Lee, T. (2013). Google Books ruling is a huge victory for online innovation. *Washington Post*. <https://www.washingtonpost.com/blogs/the-switch/wp/2013/11/14/google-books-ruling-is-a-huge-victory-for-online-innovation/>

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182. doi:10.1126/science.1199644 PMID:21163965

Nunberg, G. (2009a). Google Books: The Metadata Mess. *Google Book Settlement Conference*. UC Berkeley. <http://people.ischool.berkeley.edu/~nunberg/GBook/GoogBookMetadataSh.pdf>

Nunberg, G. (2009b). Google's Books search: A disaster for scholars. *Chronicle of Higher Education*. <http://chronicle.com/article/Googles-Book-Search-A/48245/>

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS One*, 10(10), e0137041. doi:10.1371/journal.pone.0137041 PMID:26445406

Sato, Y. (2010). Keio gijuku daigaku ni okeru Google-Library-Project no chosakuken chosa nitsuite. *Media Net*, 17, 51-53. <https://www.lib.keio.ac.jp/publication/medianet/article/017/01700500.html>

Smith, E. (2013). Google Books scores big legal victory, authors [sic] guild plans appeal. *Associations Now*. <https://associationsnow.com/2013/11/google-books-scores-big-legal-victory-authors-guild-plans-appeal/>

Texas Digital Library (TDL). (2019). *Home page*. <https://www.tdl.org/>

Thylstrup, N. B. (2014). *The Politics of Mass Digitization*. Københavns Universitet, Det Humanistiske Fakultet.

U.S. Census Bureau. (2011). Language Use. *U.S. Census Bureau*. <https://www.census.gov/hhes/socdemo/language/>

Venkatraman, A. (2009). Europe must seize lead on digital books, urges reading. *Information World Review*, (260), 1-1. <https://search.proquest.com/docview/199357636?accountid=7285>

VLDL. (2011). *Fourth Workshop on Very Large Digital Libraries*. <http://www.delos.info/vldl2011/>

Weiss, A. (2016). Examining Massive Digital Libraries (MDLs) and their impact on reference services. *The Reference Librarian*, 57(4), 286–306. doi:10.1080/02763877.2016.1145614

Weiss, A. (2018). *Big data shocks: an introduction to big data for librarians and information professionals*. Rowman & Littlefield.

Weiss, A., & James, R. (2013a). Assessing the coverage of Hawaiian and Pacific books in the Google Books digitization project. *OCLC Systems & Services: International Digital Library Perspectives*, 29(1), 13–21. doi:10.1108/10650751311294519

Weiss, A., & James, R. (2013b). An examination of Massive Digital Libraries' coverage of Spanish language materials: Issues of multi-lingual accessibility in a decentralized, mass-digitized world. Kyoto, Japan. *2013 International Conference on Culture and Computing*, 10–14. 10.1109/CultureComputing.2013.10

Weiss, A., & James, R. (2014). *Using Massive digital libraries: A LITA Guide*. Neal-Schuman Publishing.

Weiss, A., & James, R. (2015). Comparing the access to and legibility of Japanese language texts in Massive Digital Libraries. *2015 International Conference on Culture and Computing*, 57-63. 10.1109/Culture.and.Computing.2015.51

Zuboff, S. (2015). Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30(1), 75–89. doi:10.1057/jit.2015.5

ADDITIONAL READING

Bellardo, H. (2008). Mass digitization. Implications for preserving the scholarly record. *Library Resources & Technical Services*, 52(1), 18–26.

Chen, X. (2012). Google books assessments: Comparing Google Books content with WorldCat Content. *Online Information Review*, 36(4), 507–516. doi:10.1108/14684521211254031

Ching-man, A., & Jatowt, A. (2011). Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*, Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, New York, NY, USA, 1231-1240. Doi: 10.1145/2063576.2063755

Grogg, J., & Ashmore, B. (2008). Google book search libraries and their digital copies. *Journal of Library Administration*, 47(1-2), 125–140. doi:10.1080/01930820802110993

McEathron, S. (2010). An assessment of image quality in geology works from the HathiTrust Digital library. *Geological Society of America Abstracts with Programs*, 42(5), 399.

Miller, W., & Pellen, R. M. (2009). *Googlization of libraries*. Routledge.

Schmitz, D. (2008). *The Seamless Cyberinfrastructure, The challenges of studying users of mass digitization and institutional repositories*. Washington D.C.: Council on Library and Information Resources. www.clir.org/pubs/archives/schmitz.pdf

U.S. National Commission on Libraries and Information Science (NCLIS). (2006). “Mass Digitization: Implications for information policy,” Report from “Scholarship and libraries in transition, A dialogue about the impacts of mass digitization projects, March 9, 2006”. <https://permanent.access.gpo.gov/lps86366/MassDigitizationSymposium-Report.pdf>

Vaidhyanathan, S. (2011). *The Googlization of Everything: (And Why We Should Worry)*. University of California Press.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs.

KEY TERMS AND DEFINITIONS

Big Data: The large-scale aggregation of data from software and wired devices, used for the purposes of analyzing systems or user behavior; it is characterized by its large volume, wide variety, velocity (or speed of transfer), its increased variability, and undetermined veracity.

Digital Corpus: The full text of millions of books digitized by the MDLs. These provide opportunities for scholars to examine the frequency of terms as they appear in the print corpus.

Digital Humanities (DH): A branch of the humanities incorporating digital search, digitized texts, encoding, and other strategies to move previously print-bound materials and scholarship into computer science related analysis.

Fair Use: Defined by some as the breathing room to allow for the freedom of expression, it is a strong limitation of the extent of copyright law. It is determined by examining four factors: 1) the purpose and character of the use; 2) the nature of the copyrighted work; 3) the amount and substantiality of the portion taken; and 4) the effect of the use upon the potential market.

Mass-Digitization: The practice of quickly and thoroughly digitizing items on a large scale. In the case of MDLs these efforts involve scanning millions of books with multiple institutional partners across national boundaries.

Massive Digital Libraries (MDLs): Term adopted to describe the mass-digitization of printed books and the mass-aggregation of their metadata into online, full-text-searchable digital collections; some component of open access and use of public domain works defines their collections; diffuse and diverse target end-user groups are also characteristic of MDLs.

Ngram Viewer: A tool developed by Google to aid in the visualization of the digital corpus of books. It plots frequencies of terms across a graph, helping users understand how common a word or concept was in the corpus at a specific moment in history.

Public Domain: Works that are no longer under copyright protection. In the United States these works tend to be those published before January 1, 1924 and unpublished works created prior to 1899. U.S. Government publications are in the public domain.

Surveillance Capitalism: A term coined by Shoshanna Zuboff to describe practices of monetizing data derived by companies tracking online users of various services and devices.

Section 15

Medical Ethics and Law

Protocol Development in Clinical Trials for Healthcare Management

15

Swati Changdeo Jagdale*School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India***Asawaree Anand Hable***MAEER's Maharashtra Institute of Pharmacy, Pune, India***Anuruddha R. Chabukswar***School of Pharmacy, Dr. Vishwanath Karad MIT World Peace University, Pune, India*

INTRODUCTION

Clinical research is a branch of medical science which deals with any research or study in living humans. Clinical research is a process consisting of pre-clinical testing of newly developed drugs and filing of Investigational New Drug Application (IND). Phases in clinical trial involves three phases. Phase I is for assessment of safety of drug. Phase II involves effectiveness testing. Phase III is for large-scale testing, approval to use new drug i.e. licensing, approval of drug and post-marketing surveillance for reporting of any other adverse events. (Manavalan and Sinfield, 2017; Thorat, et. al., 2010; Gandhi, 2011)

Clinical research involves systematic investigation for treating the diseased conditions, overall health improvement and longevity of human. This involves volunteer participants for biomedical research with scientific guidelines. The research is conducted with pre-defined procedures, investigations are carried out carefully and outcome is recorded. The ethical principles are strongly utilized for protection of identity of participants. The clinical research allows identifying new curative and preventive ways to understand the prognosis of disease, diagnose and treat the diseases in human. (Thorat, et. al., 2010; Gandhi, 2011; Bajpai, 2013).

CLINICAL RESEARCH

Clinical trial is a research study, also called as clinical study. This study includes testing of new treatments or modification of existing therapy on human volunteers. The study determines the safety and efficacy of drug therapy for commercial use in various populations. They provide better way to screen, diagnose, prevent and cure a disease. Clinical trials are carried out for the new treatments under research whose outcome is unknown and previously not studied. Clinical trials follow a pre-designed plan for health-related research studies in human. Clinical trial is defined as the systematic study of new drugs in human to generate data for discovering and verifying the clinical, pharmacological and adverse effects with objective of determination of safety and efficacy of the new treatment. (Bajpai, 2013; Manavalan and Sinfield, 2017)

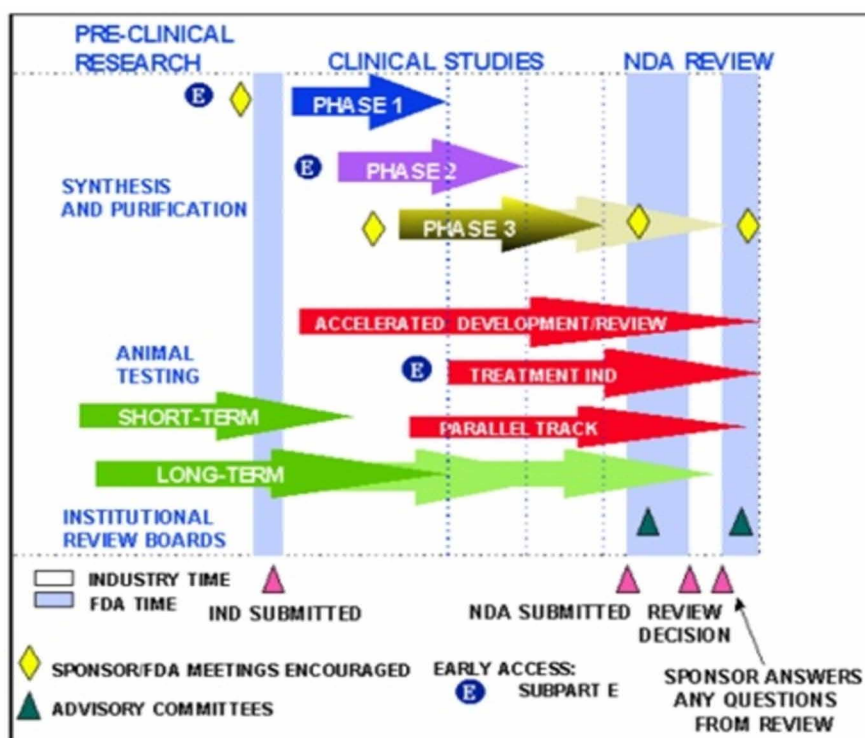
Research team in clinical trial includes doctors, nurses and health care professionals. They monitor participants and do their health checkups regularly. The research team works with participants in all

DOI: 10.4018/978-1-7998-3479-3.ch124

types of trials. Role of participants or volunteers is more active in clinical research. They are important as they get new therapies to treat a disease before these are available for commercial use. For the safety of volunteers or participants involved in clinical research, there are guidelines for appropriate selection. After qualifying the eligibility criteria they can take part in clinical trial. (Crisp and Burke, 2005; Gandhi, 2011).

The plan with all details of the clinical trial is called as research protocol. Every study has different protocol depending upon type of trial. If the protocol is carefully followed during the clinical trial process then trial becomes more successful. Design of trial plan takes care of safety of participants. So, the research protocol should be designed carefully. The protocol describes all the details of the study. (Thorat, et. al., 2010). Figure 1 represents general flow for study of the drug throughout clinical research.

Figure 1. Representation of clinical research and phases of clinical trial (Source: Thorat, et. al, 2010)



CLINICAL TRIAL PHASES

Preclinical trial study is done on laboratory animals before testing the drug on human. This testing determines safety of drug for human consumption in disease treatment. After completion of preclinical study, filing of an investigational new drug application (IND) must be done. (Thorat, et. al., 2010; Gandhi, 2011).

Phase I study is carried out on 20-80 healthy human volunteers. It determines the safety, pharmacological actions including adverse effects, drug dose and tolerability. The drugs used in the treatment of life-threatening diseases are not studied in this phase. Phase II is carried out on 100-300 patients meeting selection criteria called as volunteers or participants. It is aimed for safety, efficacy and effectiveness of the drug and determines the therapeutic potential of the drug. It also reports the adverse effects of

the drug. Phase III is carried out on around 1000-3000 participants. It is aimed to investigate safety and efficacy of drug at large number of population. Once it is confirmed that the research drug is safe and effective, a New Drug Application (NDA) is filed for approval of use of drug. The regulatory authority reviews the application and approves the drug. After approval, the drug is marketed and healthcare professionals can prescribe it. The regulatory authorities can review the drug effects at any point of time to ensure safety and effectiveness of the drug, after it is released in market for variety of population. These are Phase IV clinical trials which are also called as post-marketing surveillance. Thus, clinical trials ensure that a new drug or device or treatment is beneficial for the patients involved in the research for the diagnosis and treatment of the disease. (Crisp and Burke, 2005; Manavalan and Sinfield, 2017; Silverman and Kwiatkowski, 1998 ; Sipido, 2014).

If the participants or volunteers want to take part in clinical trials, they can directly search ongoing clinical trials on ClinicalTrials.gov by use of a registry channeled by the U.S. National Institutes of Health and National Library of Medicine. The clinical trials in India are regulated by the drug & cosmetics rules, 1945. Schedule Y along with rules 122A, 122B, 122D, 122DA, 122DAC and 122E governs clinical trials in India. (Mondal and Abrol, 2015). The Ministry of Health and Family Welfare (MoH and FW) controls Central Drugs Standard Control Organization (CDSCO) operations. CDSCO specifies the standards for clinical trials in India which ensure the safety and efficacy of trial drugs, diagnostics and medical devices. ClinicalTrials.gov currently lists 312,117 studies with locations in all 50 States and in 209 countries. (CDSCO- Amended Version -2005; Saxena and Saxena, 2014 ; Gogtay, 2017)

Currently, India is conducting 3,758 clinical trials. The registration of clinical trials carried out in Clinical Trials Registry of India (CTRI). Total number of clinical trials registered in India increased from 859 in 2015 to 3869 in 2018. (Dieppe, et. al. 2004). Figure 2 represents increasing number of registered clinical trials worldwide from the year 2000. Table 1 summarizes the percentage of registered studies location wise (as of July 25, 2019) for total number of 312,117 clinical trial. Figure 3A pie chart is a graphical representation of the distribution of clinical trials by location which is registered on ClinicalTrials.gov. ((Mondal, & Abrol, 2015, Sengupta, 2009, Lahiry, et. al., 2018).

Figure 2. Total number of studies posted since 2000 (Source: <https://clinicaltrials.gov>)

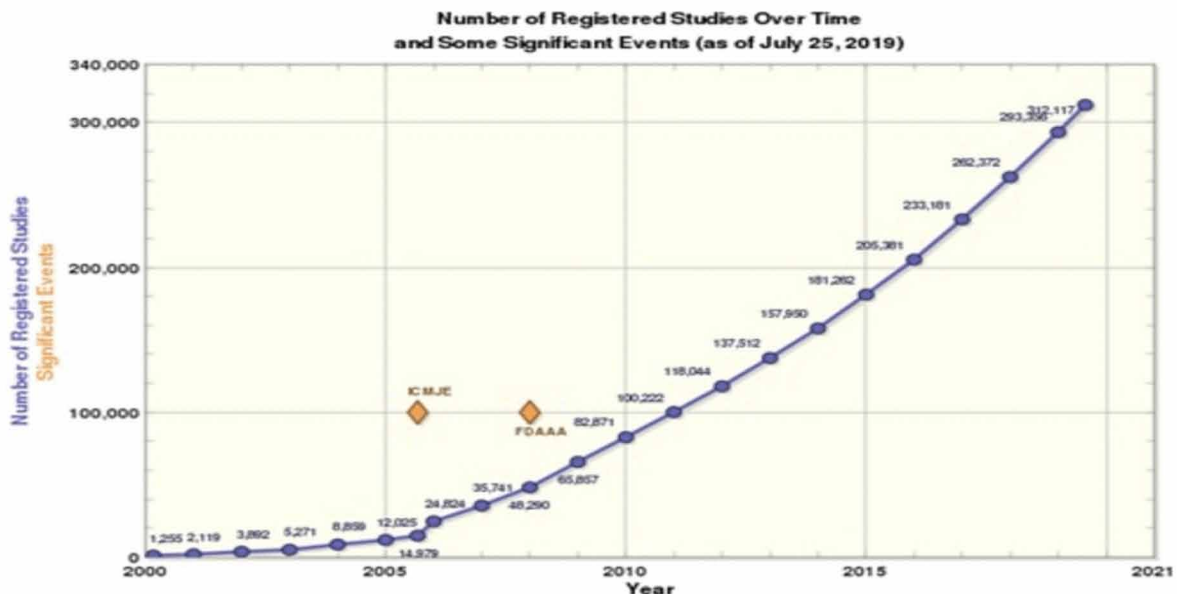
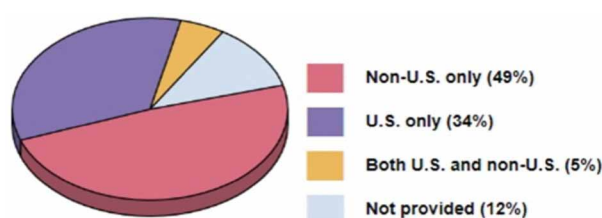


Table 1. Percentage of registered studies location wise

Location	Number of Registered Studies and Percentage of Total
Non-U.S. only	151,622 (49%)
U.S. only	106,874 (34%)
U.S. and non-U.S.	16,453 (5%)
Not provided	37,168 (12%)
Total	312,117 (100%)

(Source: <https://clinicaltrials.gov>)

Figure 3. Percentage of registered studies location wise (Source: <https://clinicaltrials.gov>)



Companies like Syneos Health, IQVIA, Paraxel International Corporation and PRA Health science are leading in clinical research sector. (Bhave and Menon, 2017). Table 2 summarizes some Indian companies with research drugs which conducted clinical trials with a positive research outcome for a specific indication with successful U.S. FDA approval and marketing of drugs. (Manavalan and Sinfield, 2017; USFDA, 2010)

Table 2. List of companies with study drugs and Indication approved by US FDA (Source: Manavalan and Sinfield, 2017).

Company	Compound Researched	Indications
Cangene	HepaGam	Hepatitis B
Eli Lilly	Alimta	Cancer
Eli Lilly	Cialis	Erectile dysfunction
Eli Lilly	Xigris	Septicemia
Janseen	Risperdal	Psychosis
Novartis	Zelnorm	Irritable bowel Syndrome
Pfizer	Vfend	Fungal infections
Wyeth	FluMist	Influenza
GlascoSmithKline	Lamictal	Epilepsy
Alcon	Vigamox	Ophthalmic infections

CLINICAL TRIAL STANDARDIZATION

For standardization in the study process there are several documents ensuring that all the staff is working in the same way. These documents have prime importance in clinical research especially if the trial is multisite. The standardized trial is also considered as ethical study with valid data. The documents like investigator's brochure, operations manual, standard operating procedures (SOPs) and research protocol ensure the study is going in standardized manner. (Sipido, 2014; Gandhi, 2011; ICMR, 2006).

The investigator's brochure contains clinical and non-clinical data derived from previous experiments. The operation manual provides detailed procedures required in the actual research. The standard operating procedures (SOPs) describe detailed specific procedures. Most SOPs are identical for all trial study sites. All staff should be trained properly for all procedures involved in the study. The research protocol describes a plan in detail for the essential aspects of the proposed study. (Gandhi, 2011).

BACKGROUND

Clinical Trial Protocol

Clinical trial is conducted as per the predefined action plan. After design, complete planning of the study, the plan should be written step wise in proper manner. This plan is called as study/ research protocol explaining the study methodologies in detail. Every clinical study has its written protocol. Depending upon the type of clinical research the complexity of protocol varies. Thus the development of protocol is an important step. If the study is multisite then the protocol should be more elaborative to avoid the risk. A protocol should be designed and written carefully for the health benefits and safety of the participants involved in the study. ((Mondal, & Abrol, 2015, Patel, et. al., 2019).

Protocol writing is a collaborative effort by a team which makes sure that the scientific procedures are practically applicable. A well written up protocol gives good study design and a guiding line for the staff involved in the research to conduct the study properly. It should clarify the thinking process of investigator and purpose of the study clearly. The success of the study trial depends on the protocol so the researchers should add every minute detail of methodology in it. (Christina, 2015, Al-JunDi, & Sakka, 2016).

There are some benefits of good protocol to investigators and other staff. Protocol acts as a detailed guide throughout the study. It allows staff to plan accordingly, the steps and review them. It helps to the researchers to adhere to timeline and budget estimates. The principal investigator is a doctor who leads the protocol designing. During the protocol development investigator can take beneficial advices from experts and other staff to refine the plan and procedures. (Bagatin, et al., 2013).

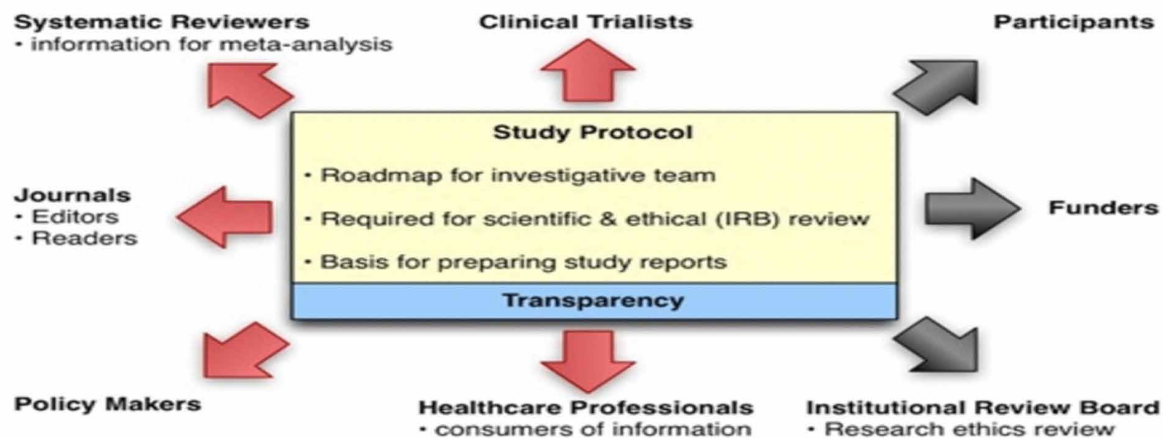
A protocol provides evidence for the prerequisite and feasibility of a study in written form. A research protocol is a document containing all the details about the background, need of study, aim, methodology, data management, ethics, time and finance of a clinical research project. The researchers should mention various components of the study including trial procedure, study monitoring, data collection methods, outcomes, statistical approaches and analysis of results etc. A protocol decides the eligibility of the participants in the study. It gives detailed information about the drug dose, frequency, tests procedures, period to which study should be carried out. The protocol information should be clear, easy to understand by another investigators and staff. It should be explained in detail to answer any question related to research. (Christina, 2015 ; Eaton and Santini, 2007).

After the completion of development and approval of the protocol, staff should strictly follow the protocol during the study. Violations of the procedures may reflect badly on the complete study data and results. The violated part of the study can be excluded during analysis of data and results. After completion of written protocol, it should be reviewed by the suitable persons. The protocols are lengthy and complex with numerous steps of scientific methods involved. To avoid any mistakes, the review of protocol should be more than once. The team of staff involved in the study should regularly monitors the medical and behavioral conditions of the volunteers in the study. This monitoring ensures the safety and effectiveness of the protocol. (O'Brien and Wright, 2002;Christina, 2015).

Figure 4 represents the importance of clinical study protocols and the associated beneficiaries.

Figure 4. Importance of clinical study protocol

(Source: <https://irb.research.chop.edu/protocol>)



PURPOSE OF A CLINICAL TRIAL PROPOSAL (Sipido, 2014)

- It highlights the importance of the research.
- It provides objective and suggest methodology in detail to achieve the objectives.
- It clarifies the hypothesis of research.
- It highlights the benefits, limitations and demands to achieve the objectives.
- It provides detailed ethical considerations.

NECESSITY OF CLINICAL TRIAL PROTOCOL

The clinical trial should be novel and ethical. Successful clinical trial run depends on the relevancy of trial to its objective, interest of the investigator and staff to adhere to the guidelines, feasibility to carry out all steps without error. It provides transparency about the procedures to everyone involved in the clinical study. (Christina, 2015; Al-JunDi, and Sakka, 2016).

Incomplete protocol leads to ambiguous procedures and inconsistent results, thus many problems may arrive during research trial. For high quality clinical research studies a complete and well written protocol play a vital role. A carefully planned and well written protocol helps to conduct and complete

clinical trial smoothly. To avoid any issues during the trial run, it is necessary to spend some time initially to write a detailed protocol before the trial begins. (Eaton and Santini, 2007).

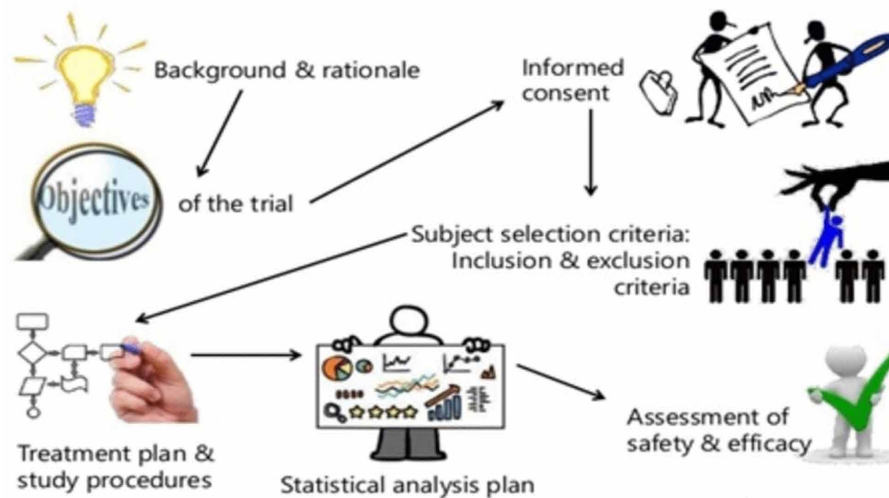
WRITING A GOOD CLINICAL TRIAL PROTOCOL

A good research protocol should clearly highlight the aim and objective of the study which helps investigator to stick to the purpose. It should answer every question raised by the investigator and other staff involved. It should explain the procedure of every step in detail, involved in the research to carry out the trial. This will avoid any confusion and errors during the trial. The scientific challenges, practical applicability of the procedures, regulatory guidelines and reporting timelines should be addressed in detail.

Detailed literature survey is must to define a research problem and to formulate aim and objective of the study. It should clearly define the study population, eligibility criteria for the participants and research outcome. Sample size and statistical calculations should be defined. Timelines for review, monitoring, audits and reporting should be predefined. (O'Brien and Wright, 2002 ;Christina, 2015;Hulley, et. al. 2001)

After completion of writing of protocol and before starting the research study, the designated Institutional Review Board (IRB) must approve the protocol. There is requirement of clinical trial approved protocol for fulfillment of ICH-GCP guidelines. Title 21 Part 312 of the Code of Federal Regulations (21 CFR 312) narrate that a research protocol and its amendments for clinical studies are managed under an Investigational New Drug (IND) application.(ICH guidelines, 2016; O'Brien and Wright, 2002;Christina, 2015). Figure 5 represents the steps involved in clinical trial protocols.

Figure 5. Components of clinical research protocols



CLINICAL TRIAL PROTOCOL CONTENT

(Bagatin, et al., 2013; World Health Organization., 2014;Rodger, 2015;ICH guidelines, 2016; Al-JunDi and Sakka, 2016; O'Brien and Wright, 2002; Christina, 2015).

A protocol should explain the following points in details as per ICH- Good Clinical Practice (GCP):

- Title Page (General Information)
- List of Abbreviations
- Background Information
- Objectives/Purpose
- Study Design
- Selection of Subjects
- Treatment of Subjects
- Assessment of Efficacy
- Assessment of Safety
- Adverse Events
- Direct Access to Source Data/Documents
- Data Handling and Recordkeeping
- Data Management
- Statistics
- Follow-Up
- Strengths and limitations
- Quality Control and Quality Assurance
- Expected Outcomes of the Study
- Ethics
- Budget
- Publication Policy
- Project Management
- Conflict of Interest
- Timeline
- References
- Annexure
- Supplements

DETAILED CONTENT OF CLINICAL STUDY/RESEARCH PROTOCOL

Title Page and General Information

Title page and general information section should include the content summary with administrative details of the protocol. The title of the study trial should be informative, concise and reflect the study details. This section should include following details:

- Title of research protocol with protocol number
- List of content: divided in section and sub-sections accordingly
- Name, address and signature of the authorized team member with date, approved version number of protocol
- Name and address of the sponsor or funding agency
- Name, title address and telephone number(s) of the investigator(s) with responsibilities

- Name, title, address, and telephone number(s) of the qualified physician
- Name, title, address, and telephone number(s) of all the staff or team members involved in trial
- Name, address and telephone number(s) of the research site(s)
- Name(s) and address(es) of the clinical laboratory(ies) and other technical service providers

List of Abbreviations

This section should enlist with the abbreviation used in the study protocol.

Background Information

This section gives detailed information about the scientific reason for the research problem for a particular protocol. The scientific literature survey and justification of the selected research should be mentioned in this section. Introduction should be in a sequential manner. It should contain a name and description of the investigational product(s). It should also contain a summary of the research protocol. It should be clear, concise, includes a significance of research and all important aspects of the protocol like aim and objectives, number of participants, methodology, timelines and expected outcome. The literature survey should state the rationale of the research. The aim and objective of the protocol should be in continuation with the significance of the research mentioned in the literature review section. It should give the answers to all possible questions related to selected clinical research.

Objectives / Purpose

This section includes the aim and objectives of the selected clinical research problem. The description of the objectives and purpose of clinical research should be simple, clear, detailed, easy to understand, specific to research problem. The purpose and objectives of the research protocol should be logical, sequential, realistic, relevant to the research problem emerged from the literature review. It should be feasible, time based and not to be over ambitious. Major and minor objectives should be described in this section of protocol.

Study Design

This is most important part of protocol. This section describes the design of the study and detailed procedures involved in research trial in stepwise manner. This should include detailed information regarding study site, participants, detailed methodology, sample size and observations. The study design should be in relation to the objectives of the study trial. The inclusion and exclusion criteria for eligibility of participants depend on the type of the study and should be defined properly.

The details about the observations to be checked should be mention in this section with techniques and procedures to note the observations, the time period and frequency to which the observations need to be made. If the observations are in format of questionnaire, then it should be attached to the protocol in the end. All laboratory or diagnostic test procedures should be mentioned in detail. (Noordzij, et. al. 2009)

The sample size of the research is the indicative of the target/ selected population and should be previously defined and justified. The basis on which sample size has been decided should be included in this section. A larger sample size increases the time required for the study completion and cost of the study. The utilization of large number of subjects to test the research hypothesis and conduct the study trial is

not ethical. The smaller sample size can be studied within short period of time. It should be economical and ethical. Small sample size also can be unethical if exposed to risk without understanding of science.

Selection of Subjects

This section describes the selection and withdrawal of subjects or participants in study trial i.e. the inclusion and exclusion criteria for participants. This is important part of protocol as it provides information on eligibility criteria for the subjects. This should highlight all the appropriate selection criteria like age, sex, type of underlying disease, medical history, and other medical conditions. The eligibility criteria help in practical applicability, safety issues of participants and meeting to the regulatory guidelines. If criteria are not well written, the validity of outcome of the study trial may affect.

Eligibility Criteria Should Have:

- Number of eligibility criteria should be minimum
- Criteria should ensure the safety of the patient and correct scientific knowledge
- Criteria should be clear, well defined for easy to understand.
- Criteria should be verified by an auditor.

Failure to write eligibility criteria properly leads to failure to imitate clinical practice. It may increase complexity and costs required for the study. It may also lead to reduced number of recruitment of participants involved in study.

Treatment of Subjects

This section includes all the detailed information regarding treatment given to the participants. This contain the name(s) of drug, drug dose(s), dosing frequency (ies), the route (s) of drug administration for all the product(s) used in the treatment and follow up schedule(s) for participants for each product involved in the trial.

Assessment of Efficacy

This section provides the specifications of the parameters for efficacy study. This should describe methods for record, assessment, and analysis of efficacy parameters. Overall this highlights effectiveness of the study.

Assessment of Safety

This section describes the specifications of the parameters of safety for the study. It describes the expected side effects and adverse effects of the associated study drug. The detail reporting of adverse effects to the sponsor and regulatory authorities should be done. The follow up requirement after adverse events should be mention. This should detail the procedure to deal with and after adverse events. This should describe methods for record, assessment, and analysis of safety parameters.

Direct Access to Source Data/Documents

This section describes the permission from investigators to directly access the collected data for periodic monitoring, review, audits, and review by institutional review board (IRB) and inspections by regulatory agencies.

Data Handling and Recordkeeping

This section of protocol should provide information of methods of data collection, data handling, data management and storage. Some data collection tools are like questionnaires, structured or semi-structured interviews, medical records, results of laboratory tests, other clinical examinations etc. The data can be stored in computers for analysis. Further the collected data can be monitored, verified and reviewed to summarize in research outcome. The data coding in computer systems should be mentioned in the protocol. The statistical methods used should be predefined to carry out analysis of the collected data.

Data Management

A detailed data management plan should be mentioned in this section. A general description for data management for protocol should be provided. The methods for data collection, documentation, submission, review, verification, monitoring and archival should be explained in detail.

Statistics

This section should describe the statistical methods to be used in the trial to analyze the data. As the significance of the study trial is based on the statistical results, this section should be carefully designed. For writing these procedures a statistician should be involved. The number and name of variables used in the analytical calculations should be listed. For computerized statistical analysis methods, the name and versions of the software utilized should be mentioned. This section should also describe the statistical testing of data collected in on-going study, after completion of study and periodic review. This section should also explain and justify the statistical procedure to be followed for the deviations.

Procedure for Accounting for Unused and False Data:

This section gives knowledge about the procedure to follow to report any deviation from the pre-defined statistical plan and justification for it.

Follow-Up

This section should highlight the necessity of the follow up to the participants and for how much period of time follow up to be taken. The follow up are for any adverse events, any change in drug dose and frequency, missed dose etc. The follow up can be taken even after completion of collection of data for a specific study.

Strengths and limitations

This section should describe the benefits, advantages, limitations and disadvantages of the study. This section helps in proper resource allocation for the work.

Quality Control and Quality Assurance

This section should inform about the plan for controls and standards for confirmation of accuracy followed in each step. A protocol should provide a general description of methods for the quality control and assurance of the study.

Expected Outcomes of the Study

This section provides information about the effect and contribution of the study to the current knowledge. The utilization of research outcome for health benefits are mentioned in this section. It also helps information of new healthcare policies.

Ethics

The protocol should describe ethics and confidentiality of the study in relation to protection of participants. This section should mention the process of ethical approval and from whom it is to be taken. It should also address the issues which raise ethical concerns of the participants. This section should describe about the informed consent, plan and process to take consent from the participants. There should be freedom to the participants to withdraw from the study at any point of time as per their wish. The safety and secrecy for data should be maintained.

Informed Consent Forms

The informed consent form format must be there in approved protocol. The consent form have be approved by Ethical Review Committee (ERC).The consent form should be in two language i.e. English and local language. The consent form should explain the risk and benefits associated with the trial, justification of the research purpose with social and physical implications.

Budget

This section describes the finances and insurances of the study in detail. In some research protocols, it is documented in a separate agreement. This section should contain a detailed item-wise breakdown of the funds requested, with a justification for each. It should also consists of head wise funds received and expenditure in detail like for staff, instruments, laboratory tests, cost of the drug, transportation etc. The estimate of finance is to be attached in annexure. All costs including personnel, consumables, equipment, supplies, communication, and funds for patients and data processing are all included in the budget. Each item should be justified.

Publication Policy

This section provides the publication policies, guidelines and procedures of study findings publication in detail. The publication policy should be discussed regarding the leads and acknowledgements in the publication. Investigator should know and follow the publication requirements and policies from institution and sponsor. In accordance with the Food and Drug Administration (FDA) Amendments Act (42 CFR Part 11), clinical trial findings and results will also be published on a public website, ClinicalTrials.gov. Sometimes publication policy is documented in a separate agreement.

Project Management

This section should describe the role and responsibility of each person of the team involved in the study.

Conflict of Interest

Identify and document clearly; any consultative relationship that the principal investigator or co-investigators found unrelated to the protocol that might be considered a real or apparent conflict of interest.

Timeline

This section should give the timelines of starting and completion the study. The protocol should specify the time for each phase of the project. The proper time table for accomplishing each activity of the study should be defined. The protocol should specify the time along with a detailed month by month timeline for each activity. Assigning time frame to each step in the trial will be helpful in organizing the structure of the research trial. The adherence to the time schedule will allow study completion on time.

References

Plagiarism is copying other's research and considered as a crime. Plagiarism is failure to referencing and acknowledgement to other researcher's work. So, referencing is an extremely important aspect of the research protocol.

Referencing is the recognizing the information taken from the other research work. A citation for the background information will be useful for the readers to take reference for an interested area. All references should be listed at the end of the protocol. The reference formatting depends upon the funding agencies to where submission of protocol will be done. In clinical writing, the most commonly used citation systems are the Vancouver and the Harvard system. **Annexure**

The annexes are need to be attached at the end of the protocol:

1. Informed consent form.
2. Letters from ethics committees.
3. Study questionnaire (copies of any questionnaires or draft questionnaires).
4. Case Record Forms (CRFs).
5. Budget details.
6. Curriculum Vitae (CV) of the chief investigator and co-investigator and their role in the study.

Supplements

Any additional information regarding research is provided in this section. The templates of informed consent, forms, manuals, handbooks are attached in this section.

GOOD PROTOCOL

Good protocol ensures well-defined role of every personnel involved in study. Study should yield the expected outcome of research. The protocol should answer the research questions adequately. The protocol should give detailed methodology, which can be guidance document for another investigator also. (Al-JunDi, and Sakka, 2016)

There are some common mistakes to avoid during writing of protocol. Insufficient parameters, inadequate explanation, sample size, excess of work in less time period, practical applicability of the procedures, unreliable instrumentation, improper statistics are need to take into considerations and analyzed. (Eaton and Santini, 2007).

As research protocols are lengthy, there might be chances of human errors. Use of some tools for any spelling mistakes and grammatical mistakes can be carried out to avoid it. The protocol can be checked by using automatic software, peer reviewing and proofreading to avoid mistakes. (O'Brien and Wright, 2002)

The templates of clinical research protocols are available for guidance and reference. The template varies depending on the phase of clinical trial, type of design of study like intervention studies, descriptive, observational and registry studies. The protocol writing guidelines and templates are freely available on <https://clinicaltrials.gov>, <https://grants.nih.gov/grants/oer.htm>, https://hub.ucsf.edu/protocol-developmentandhttps://www.who.int/rpc/research_ethics/format_rp/en/ etc.

PROTOCOL REVIEW

After finishing complete write-up of the protocol, it should be reviewed and approved by the experts. The approval and monitoring of clinical trial must be done by an Institutional Review Board (IRB). This is an independent committee which includes healthcare professionals like physicians, dentists for review of scientific information, statisticians for data analysis and members of the society as a common man. The committee ensures the safety of the participants and ethics of the study with the protection of participants. After protocol approval the IRB should review the study periodically. The review ensures that the protocol has minimum risk with more potential health benefits to the participants. (O'Brien and Wright, 2002 ;Hulley, et. al. 2001)

PROTOCOL CLARIFICATION

A protocol clarification helps into the implementation and conduct for the research study. It clarifies no any change in protocol. It also gives internal guidance. There is no need of submission of clarification to IRB. It should be provided in writing to all investigators. (Al-JunDi and Sakka, 2016)

PROTOCOL VIOLATION

A protocol violation also referred as protocol “deviations.” It occurs when a staff performs any action that does not adhere to the research protocol. It may be an omission, addition or change in described procedure. The violation of protocol is a human error and may be because of a problem with study oversight, training to personnel, or study procedures. The study personnel should take efforts to stick to the protocol and not to make violations. If it occurs, then corrective action should be taken, written, notified and submitted to the IRB. The violations must be informed and reported to the study investigator, sponsor and regulatory authorities if required. If violations are repetitive by the staff, there is need of training or a protocol amendment depending upon the type of violations. (O’Brien and Wright, 2002; Al-JunDi and Sakka, 2016).

PROTOCOL AMENDMENT

If there is need to update the research protocol, then the changes should be described in detail in written document. Before implementation, the amendments should be submitted in written to the Institutional Review Board for review and approval. But if the changes are administrative type or immediate removal of any hazards to participants then there is no need of prior approval from IRB. The amendments are need to submit to the sponsor. If the study involves a product regulated by any specific regulatory authority, the amendment must be submitted to the same and also to IRB before implementation. After approval of protocol amendment, all the research participants must be informed about the changes. If required, the informed consent should be also revised and signed by the participants as per the changes. (Al-JunDi and Sakka, 2016; Hulley, et. al. 2001, O’Brien and Wright, 2002).

FUTURE RESEARCH DIRECTIONS

A good protocol can be a reference guide for other clinical trial investigators. The protocol development in clinical trial should be more focused and carefully written. This document will be providing a self explanatory detailed guideline for various elements and procedures to be followed during the clinical trial study. Also the justification of all the processes, sample size selection, lab tests, statistical analysis etc should be provided in the protocol. The protocol should include all instructions, clarifications and related examples in it. There is more need to pay attention to include all data related to preclinical and clinical study in protocol. So that the protocol should be answer all the questions.

CONCLUSION

Clinical research is a systematic investigation for treating the diseased conditions, overall health improvement and longevity of human. Clinical trials follow a pre-designed study plan for health-related research studies in human. This plan is called as ‘research protocol’ and is very important aspect in conducting clinical trial. To perform clinical trial properly understanding of the steps involved in the trial is essential. For better and sequential understanding of clinical trial process, methodologies and for the reliable data collection, a good protocol should be written. Protocol writing is a collaborative effort by a team

which makes sure that the scientific procedures are practically applicable. It should be accurate, clear, comprehensive, without errors and easy to understand. It should summarize all the details of the research trial. The protocol writing is lengthy and time consuming process. But the time spent to write a good protocol is worth as it saves the time and failures coming in actual conducting a clinical trial later on. Good clinical trial opens up new therapeutic options for a specific disease. The good study determines the safety and efficacy of the new treatments and helps in overall improvement of human health.

REFERENCES

- Al-JunDi, A., & Sakka, S. (2016). Protocol Writing in Clinical Research. *Journal of Clinical and Diagnostic Research: JCDR*, 10(11), 10–13. PMID:28050522
- Bajpai, V. (2013). Rise of Clinical Trials Industry in India: An Analysis. *ISRN Public Health. Article ID, 167059*, 17.
- Bhave, A., & Menon, S. (2017). Regulatory environment for clinical research: Recent past and expected future. *Perspectives in Clinical Research*, 8(1), 11–16. doi:10.4103/2229-3485.198551 PMID:28194332
- Central Drugs Standard Control Organization, Directorate General of Health Services, Government of India: Good Clinical Practices for Clinical Research in India. Schedule Y (Amended Version -2005).
- Christina. (2015). *Protocol writing in Clinical Research*. James Lind Institute (JLI).
- Crisp, R.J., & Burke, F.J. (2005). Conducting trials. *British Dental Journal*, 198(5), 282–283. doi:10.1038/bdj.4812157 PMID:15870753
- Dieppe, P. A., Ebrahim, S., Martin, R. M., & Juni, P. (2004). Lessons from the Withdrawal of Rofecoxib. *British Medical Journal*, 329(7471), 867–868. doi:10.1136/bmj.329.7471.867 PMID:15485938
- Eaton, K. A., & Santini, A. (2007). *Designing a protocol*. London. Faculty of General Dental Practice.
- EdiléiaBagatin, Helio, A., & Miot. (2013). How to design and write a clinical research protocol in Cosmetic Dermatology. *Journal Brazilian Annals of Dermatology*, 88(1), 69–75. PMID:23539006
- U. S. Food and Drug Administration (FDA) Guidelines for Conduct of Clinical trials. (2010).
- Gandhi, P. (2011). Clinical Research Methodology. *Indian Journal of Pharmaceutical Education and Research*, 45(2), 199–209.
- Gogtay, N., Ravi, R., & Thatte, U. M. (2017). Regulatory requirements for clinical trials in India: What academicians need to know? *Indian Journal of Anaesthesia*, 61(3), 192–199. PMID:28405032
- Hulley, S., Cummings, S., Browner, W., Grady, D., Hearst, N., & Newman, T. (2001). *Designing clinical research: an epidemiologic approach*. Lippincott.
- Indian Council of Medical Research: Ethical Guidelines for Biomedical Research on Human Participants. (2006). Author.
- International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline: Guideline for Good Clinical Practice E6 (R1).

Lahiry, S., Sinha, R., Choudhury, S., Mukherjee, A., & Chatterjee, S. (2018). Paradigm shift in clinical trial regulations in India. *Indian Journal of Rheumatology*, 13(1), 51–55. doi:10.4103/injr.injr_110_17

Manavalan, S., & Sinfield, C. (2017). *Conducting Clinical Trials in India: Opportunities and Challenges*. Academic Press.

Mondal, S., & Abrol, D. (2015). *Clinical Trials Industry in India: A Systematic Review*. Institute for Studies in Industrial Development, working paper 179.

Noordzij, M., Dekker, F. W., Zoccali, C., & Jager, K. J. (2009). Study designs in clinical research. *Nephron. Clinical Practice*, 113(c), 218–221. doi:10.1159/000235610 PMID:19690439

O'Brien, K., & Wright, J. (2002). How to write a protocol. *Journal of Orthodontics*, 29(1), 58–61. doi:10.1093/ortho/29.1.58 PMID:11907312

Patel, B. B., Tiwari, S., & Mishra, B. (2009). Clinical research in India: An update. *Indian Journal of Pharmacy Practice*, 2(2), 14–24.

Rodger, J., A., (2015). Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining. *Patient Informatics Processing Software Hybrid Hadoop Hive, Informatics in Medicine Unlocked*, (1), 17–26.

Saxena, P., & Saxena, R. (2014). Clinical trials: Changing regulations in India. *Indian Journal of Community Medicine*, 39(4), 197–202. doi:10.4103/0970-0218.143018 PMID:25364141

Sengupta, A. (2009). Fatal trials: Clinical trials are killing people. *Indian Journal of Medical Ethics*, 3(3), 118–119. doi:10.20529/IJME.2009.041 PMID:19653585

Silverman, R., & Kwiatkowski, T. (1998). Research fundamentals: III. Elements of a research protocol for clinical trials. *Academic Emergency Medicine*, 5(12), 18–23. doi:10.1111/j.1553-2712.1998.tb02698.x PMID:9864136

Sipido, K. R., Casadei, B., Holvoet, P., Janssens, S., Luttun, A., & Sampaolesi, M. (2014). Bedside to bench: A look at experimental research with a clinical trial checklist. *Cardiovascular Research*, 101(1), 1–3. doi:10.1093/cvr/cvt272 PMID:24357016

Thorat, S. B., Banarjee, S. K., Gaikwad, D. D., Jadhav, S. L., & Thorat, R. M. (2010). Clinical Trial: A Review. *International Journal of Pharmaceutical Sciences Review and Research*, 1(2). Article, 019, 101–106.

World Health Organization. (2014). *Guide for writing a Research Protocol for research involving human participation*. WHO.

KEY TERMS AND DEFINITIONS

Clinical Research: This research is carried out to ascertain the safety, efficacy and effectiveness of the new drug product or new therapy for treatment of a specific disease in human.

Clinical Trial: This research is experiments carried out in human, observations are recorded to determine the safety and efficacy of the therapy.

Investigator: The individuals who actually carry out and monitor the clinical trial.

Pre-Clinical Study: This is part of clinical research in which the safety of the new drug or treatment is carried out on experimental animals.

Protocol Deviations: The non-compliance to the protocol is protocol deviations but this does not significantly affect the risks and benefits of the clinical trial.

Sponsor: The sponsor can be a person or organization which provides all financial assistance to carry out a clinical trial.

Study Protocol: This document provides details of each step to be carried out in clinical research.

A Regulatory and Safety Perspective on Medical Devices

15

Carla Pires <https://orcid.org/0000-0003-4526-0271>

CBIOS, Universidade Lusófona, Portugal

Dinah Duarte

CBIOS, Universidade Lusófona, Portugal

INTRODUCTION

In Europe, medical devices market account for sales of EUR 100 billion, with the medical devices industry employing 575,000 people in the European Union (EU). This industry is one of the most representative employers in this economic area (European Commission, 2018). In 2015, the United States medical device market was evaluated at more than \$140 billion, which represented nearly 45% of the global market (International Trade Administration, 2016). By 2021, the market of medical devices is expected to globally attain \$342.9 billion, mainly due to the ageing of population (Lucintel, 2016). The main commercial trends in medical devices are, as follows: surgical and infection control, general medical devices, cardiovascular, and home healthcare. It is expected that in the future medical devices will tend to be smaller, portable, and including software (Lucintel, 2016).

The design and marketing of medical devices is strictly legislated. In accordance with World Health Organization (WHO) and Global Harmonization Task Force a medical device is defined as:

“an article, instrument, apparatus or machine that is used in the prevention, diagnosis or treatment of illness or disease, or for detecting, measuring, restoring, correcting or modifying the structure or function of the body for some health purpose. Typically, the purpose of a medical device is not achieved by pharmacological, immunological or metabolic means” (WHO, 2018; Global Harmonization Task Force, 2012).

Additionally, the Food and Drug Administration (FDA) through the 201(h) of the Federal Food Drug & Cosmetic (FD&C) Act defines a device as:

“an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is: recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them, intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or intended to affect the structure or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes” (FDA, 2018a, 2018b).

DOI: 10.4018/978-1-7998-3479-3.ch125

Particularly, medical devices are regulated in accordance to Council Directive 93/42/EEC in the EU. According this Directive, all *medical devices must be designed and manufactured in such a way, that, when used under the conditions and for the purposed intended, they will not compromise the clinical condition or the safety of patients, or the safety and health of users or, where applicable, other persons, provided that any risks which may be associated with their use constitute acceptable risks when weighed against the benefits to the patient and are compatible with a high level of protection of health and safety.* In addition, each member state of the EU has a National Competent Authority (NCA), which oversees specific designated Notified Bodies (NBs), i.e. the private entities, which are responsible for the certification of medical devices and for compliance of medical devices to EU Directives (Kramer et al., 2013; World Bank income group, 2016). NBs are responsible for evaluating the risk analysis and risk management strategy of the manufacturer (article 26 of Decree- Law No. 145/2016). In Portugal, the NCA - INFARMED, I.P. - is responsible for designating and supervising the national NBs (article 22° and 25° of Decree- Law No. 145/2016,) or for supervising the marketed medical devices (article 60° of Decree- Law No. 145/2016).

All incidents regarding medical devices (e.g. malfunction, deterioration, or performance issues) are recorded and centrally evaluated by the competent authorities of each member state (article 10 of Council Directive 93/42/EEC); if the alerts on MDs are considered relevant (e.g. potential adverse events related to medical devices), they are circulated between member states in the form of a National Competent Authority Report (NCAR) (EU statistics, 2017). Among others competences, NCAs are responsible for communicating adverse events with medical devices to the European Databank on Medical Devices (EUDAMED) (Kramer et al., 2013). EUDAMED aims at the collection of information, the exchanging of information, and conduction of marketing surveillance on medical devices in all member states (European Commission, 2001). Similarly, FDA is the authority responsible for regulating manufacturing and marketing of medical devices, since 1976. A premarket approval is usually required for higher-risk devices. The applications of these devices usually comprise clinical data and must include data to prove safety and efficacy (FDA, 2018a; Jarow & Baxley JH, 2015; Peña, et al, 2007).

Despite these strict regulations, every year safety issues involving medical devices arise. For instance, in a cohort study (2005-2010) specifically enrolling cardiovascular, orthopedic, and neurologic devices first approved in the EU, safety issues and reporting of trial outcomes were quantified in comparison with the outcomes of the approved medical devices in the United States (US). This study concluded that devices first approved in the EU were associated with an increased risk of post-marketing safety alerts: 27% (62/232) vs. 14% (11/77), respectively, for safety alerts and recalls for devices approved first in EU vs. US (unadjusted rate); 2.9 (95% confidence interval 1.4 to 6.2) for the adjusted hazard ratio (Hwang et al., 2016). Contrary to this conclusion, another study identified a small difference between absolute occurrences of serious medical device recalls between the US and EU. Considering that EU approves a higher number of medical devices than US, the proportional number of serious occurrences with these products may be even lower in the EU (Davis et al., 2011). An in-depth analysis of clinical safety and effectiveness used in health technology assessments of high-risk medical devices was conducted in Europe through the evaluation of reports published between 2010 and 2015. It concluded that the quality of scientific evidence (e.g. scientific data) is low, thus recommending the application of higher quality standards such as the obligatory implementation of methodologically robust trials (associated with other evidence sources) (Olberg, B., Fuchs, S., Panteli, D., Perleth, M., Busse, R., 2017).

BACKGROUND

The core study focus was the description and analysis of all Portuguese national safety alerts on medical devices during 2017, since these were the most recent data available at the time of this analysis.

A review of the Cochrane Library¹ using the key-words “medical device” and “alert” was performed and only four results, not related with regulatory issues, were identified. This fact confirms the need to undertake this analysis and understand deeper the potential risks associated with the use of these health products.

In this sense, it was also considered necessary to present the current classification of medical devices specifically in EU, including Portugal, as well as, to present examples of devices per their classification to enhance the intelligibility and comprehensibility of this study. Thus, study objectives were summarized, as follows:

- To present the definition and classification of medical devices in EU,
- To give some examples of medical devices per classificatory group, and
- To describe and analyze all Portuguese national safety alerts on medical devices in 2017.

METHODS

In this section, methods are presented per each one of the defined objectives.

Definition and Classification of Medical Devices in EU, Including Portugal

Definitions/classifications were extracted from the national and international guidelines, namely Portuguese Decree Law No, 145/2009 and Council Directive 93/42/EEC. In addition, the new EU Directive 2017/745 on medical devices was taking into account, although not yet transposed to the Portuguese legislation.

Examples of Medical Devices per Classificatory Group

Examples of devices per classificatory groups were extracted and translated from the official website of INFARMED, I.P., since they are validated by each of the national medicines agency, which works in collaboration with the European Medicine Agency (EMA) and other international medicine agencies such as FDA (INFARMED I.P., 2016). On the other hand, it was considered that many readers are not familiar with types/examples of medical devices.

Portuguese National Safety Alerts on Medical Devices in 2017

National safety alerts on medical devices were collected from the official website of INFARMED, I.P.² All Portuguese national safety alerts on medical devices were briefly described and analyzed for the year 2017. The qualitatively and quantitatively analysis was performed by two independent researchers.

The quantitative analysis was mostly descriptive (number and percentage of medical devices by classes).

The qualitative variables were conveniently defined based on the published text of the safety alerts, as follows: devices with (or without) identified records of commercialization at national level; devices

voluntarily withdrawal from the market; and the remaining situations were classified in the group of “other”. Finally, the types of regulatory actions taken by the national medicine agency were identified.

RESULTS

Definition and Classification of Medical Devices in EU

The concept of a medical device is not uniform in all countries of EU. In Portugal, the following definition should be considered:

a medical device means any instrument, apparatus, appliance, software, material or article used alone or in combination, including the software intended by its manufacturer to be used specifically for diagnostic or therapeutic purposes and necessary for the proper functioning of the medical device, whose main effect intended in the human body is not achieved by pharmacological, immunological or metabolic means, although its function may be supported by these means intended by the manufacturer to be used in humans to purposes: i) diagnosis, prevention, monitoring, treatment or alleviation of disease; ii) diagnosis, monitoring, treatment, mitigation or compensation for an injury or a defect; iii) study, replacement or modification of the anatomy or of a process physiological; iv) control of conception (article 3, Portuguese Decree Law No, 145/2009).

The Portuguese concept is aligned with definitions of Article 1, Council Directive 93/42/EEC, since Portugal is a member state of the EU, i.e. member states of EU are aligned concerning the definition of a medical device.

Classification of Medical Devices According to EU Regulation

Medical devices are classified in the following categories: Classes I, IIa, IIb, and III in the EU. This classification is based on rules 1 to 18 of Annex IX of Council Directive 93/42/EEC, as follows:

- **Class I** (rules 1, 2, 4-6, and 12): either does not touch patient or contact only intact skin (only the ones that are not included in any of the other classes) (rule 1); channeling or storing for eventual administration (only the ones that are not included in Class IIa) (rule 2); come into contact with injured skin; used for mechanical barrier, compression, absorb exudates (only the ones that are not included in Class IIa) (rule 4); invasive in body orifice or stoma not surgically, transient use or short term use - if only in oral cavity, in ear canal or in nasal cavity (rule 5); surgically invasive, transient use - reusable surgical instrument (rule 6) and all other active devices (not covered in other classes) (rule 12);
- **Class IIa** (rules 2-5, 8-11, 15, and 16): channeling or storing for eventual administration, for use with blood, other body fluids, organs, tissues or may be connected to an active medical device in Class IIa or higher (rule 2); modify biological or chemical composition of blood, body liquids, other liquids intended for infusion - only filtration, centrifugation or exchange of gas or heat (rule 3); come into contact with injured skin, used for mechanical barrier, compression, absorb exudates - intended to manage micro-environment of wound + others (rule 4); invasive in body orifice or

stoma, not surgically, short term use or long term use - if only in oral cavity, in ear canal or in nasal cavity and not liable to be absorbed by the mucous membrane or connected to an active medical device in Class IIa or higher (rule 5); surgically invasive long-term use and implantable devices - to be placed in teeth (rule 8); active therapeutic devices intended to administer or exchange energy (only the ones that are not included in Class IIb) (rule 9); active device for diagnosis, if intended to supply energy, to image *in vivo* distribution of radiopharmaceuticals, or for direct diagnosis or monitoring of vital physiological processes (only the ones that are not included in Class IIb) (rule 10); active devices to administer or remove medicines and other substance to or from the body (only the ones that are not included in Class IIb) (rule 11); specifically to be used for disinfecting medical devices (only the ones that are not included in Class IIb) (rule 15); and devices intended for recording of X-ray diagnostic images (rule 16);

- **Class IIb** (rules 3-7, 9-11, 14, 15, and 18): modify biological or chemical composition of blood, body liquids, other liquids intended for infusion (rule 3); come into contact with injured skin, used for mechanical barrier, compression, absorb exudates - intended for wounds which breach dermis and heal only by secondary intent (rule 4); invasive in body orifice or stoma, not surgically, long term use (rule 5); surgically invasive, transient use, - biological effect, mainly or wholly absorbed or - intended to administer medicines in a potentially hazardous manner (rule 6), surgically invasive short term use, - supply energy/ionizing radiation or - undergo chemical change in body - or administer medicines - not in teeth (rule 7); active therapeutic devices intended to administer or exchange energy, - administer or exchange energy in potentially hazardous way or - intended to control or monitor or influence directly the performance of a Class IIb active therapeutic device (rule 9); active device for diagnosis, if intended to supply energy, to image *in vivo* distribution of radiopharmaceuticals, or for direct diagnosis or monitoring of vital physiological processes, - specifically intended to monitor vital physiological parameters where variations could result in immediate danger, or - all devices emitting ionizing radiation and intended for diagnostic and therapeutic interventional radiology (rule 10); active devices to administer or remove medicines and other substances to or from the body - if this is in a potentially hazardous way (rule 11), devices used for contraception or prevention of sexually transmitted diseases (only the ones that are not included in Class III) (rule 14); specifically to be used for disinfecting medical devices, - specifically to be used for disinfecting, cleaning, rinsing or hydrating contact lenses, or - specifically to be used for disinfecting invasive devices (rule 15), and blood bags (rule 18).
- **Class III** (rules 6, 7, 8, 13, 14, and 17): surgically invasive, transient use - specifically to control/diagnose/monitor/correct a defect of heart or central circulatory system through direct contact, or - for use in direct contact with central nervous system (rule 6); surgically invasive, short term use, - for use in direct contact with central nervous system, - specifically to monitor/control/diagnose/correct defect of heart or central circulatory system - by direct contact, or - biological effect or mainly absorbed (rule 7); surgically invasive long term use and implantable device, - used in direct contact with heart or central circulatory/nervous system, - biological effect or mainly absorbed, - undergo chemical change in body - or administer medicines (nor in teeth), or - specific derogation: breast implants and hip knee and shoulder joint replacements (rule 8); devices incorporating integral medicinal substances liable to act in ancillary way on human body (rule 13), devices used for contraception or prevention of sexually transmitted diseases, - if implantable or long term invasive (rule 14); and devices utilizing non-viable animal tissues or derivatives, not device in contact only with intact skin (rule 17).

Currently, a new Directive on medical devices (Directive 2017/745) has become into force in EU, but it only will be transposed (i.e. legally adopted) by each one of EU countries in the next years and medical devices already certified can stay on the market during the next five years (European Parliament and of the Council, 2017). After 2020 all medical devices will be certified in accordance the new Directive 2017/745, which is stricter than the current one (e.g. probiotics will no longer be available as medical devices, and many substances currently market as medical devices may be classified as dietary supplements) (Gruenwald, 2017).

Market Surveillance vs. Classification

Market surveillance on medical devices is supported by Directives 90/385/EEC (active implantable medical devices), 93/42/EEC (medical devices), and 98/79/EC (*in vitro* diagnostic medical devices). The criteria, such as the potential inherent risks in the use of the device; the possible incidents related to the characteristics and/or operation of the device; the duration of the contact of the device with the human body; or the anatomy affected by using the device, along the above mentioned classifications allow for classification regarding the type and risk, as follows:

- Low risk devices (Class I);
- Medium risk devices (Classes IIa and IIb; being the Class IIa low average risk and Class IIb high average risk);
- High risk devices (Class III) (INFARMED I.P., 2016).

Low risk devices (Class I) are only required to “*self-declare*” that they conform with the *Essential Requirements to the national Competent Authority in their country of origin*. Clinical and/or nonclinical evidence is needed to support approval of medium and high-risk devices (Classes IIa, IIb, and III) (Gail & Norman, 2016). Details on the definitions/classification of medical devices (Classes I to III) may be consulted in the background section.

Examples of Medical Devices per Classificatory Group

Some examples of medical devices by class are presented in Table 1 (INFARMED, I.P., 2016):

Overall, 32 safety alerts on medical devices were identified in Portugal during the year of 2017: 6 (18.8%) were related to devices voluntarily withdrawn from the market (e.g. counterfeit or falsified devices that were recalled from the market), 7 (21.9%) were related to other situations, and 19 (59.4%) were related to medical devices without identified records of commercialization at the national level (e.g. alerts disseminated by NCAs from United Kingdom, Denmark, and Poland) (Figure 1).

Most predominant actions taken by INFARMED, I.P. (NCA)

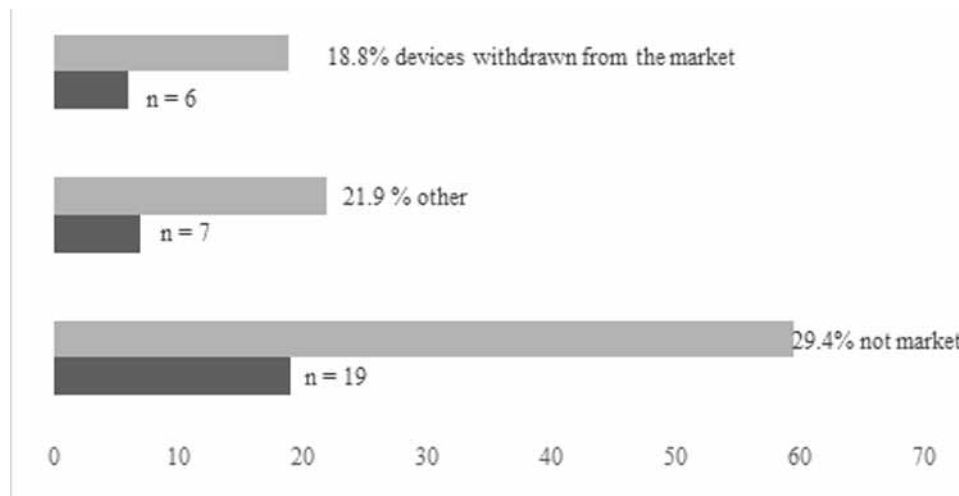
- For inexistent records or voluntary withdrawals of medical devices: the medical devices must not be used, sold, or distributed;
- For voluntary withdrawals or falsified medical devices: all devices must be returned to the manufacturer, and suspected adverse events should be reported to INFARMED, I.P.;
- For other situations: need to contact the distributor to obtain safety advice, update the software, or receive new instructions for use.

Table 1. Examples of medical devices per classification (INFARMED I.P., 2016)

Class	Examples	
Class I	Devices for collection of body fluids	bags of urine collectors ostomy bags diapers incontinence pads
	Devices intended to immobilize parts of the body and/or apply force or compression	cervical collars compression stockings elastic stockings knee braces for medical purposes
	Devices used for external support of the patient	walking aids wheelchairs hospital beds
	Non-invasive devices	stethoscope corrective glasses glass frames
	Devices intended for temporary content or with storage function	syringes without needle tablespoons specifically intended for administration of medicines
	Invasive devices with respect to body orifices for intended for temporary use	examination gloves
	Invasive devices used in the oral cavity as far as the pharynx, in the ear canal up to the ear drum, or in the nasal cavity Non-invasive devices that contact injured skin and are used as a mechanical barrier, for compression or absorption of exudates	Dressing material for nosebleeds solutions for irrigation or mechanical cleaning cotton wool bandages
Class IIa	Devices that are intended to manage the micro-environment of a wound	hydrophilic gauze, sterile or non-sterile occlusive adhesives for topical use
	Invasive devices with respect to body orifices, for use in the short term	urinary catheters vaginal pessaries
	Active devices with a measuring function	thermometer with battery or other associated source of energy
	Invasive devices with respect to body orifices, which are intended to be connected to an active medical device	heat and moisture exchangers
	Invasive devices for surgical nature, intended for temporary use	Needles syringes lancets surgical gloves
	Devices intended specifically to be used for the disinfection of medical devices	
	Devices that are intended to be used primarily in wounds that have breached the dermis substantially and extensively, and where the healing process can only be achieved by secondary intervention	dressing material for extensive and chronic ulcerated wounds
Class IIb	Devices that are intended for administration of medications	insulin pens
	Devices used for contraception and/or prevention of sexually transmitted diseases	male condoms diaphragms
	Devices intended specifically to be used for disinfecting, cleaning, washing, or hydration of contact lenses	Solutions of comfort for patients with contact lenses
	Devices incorporating a medicinal substance and constitute a single product that is not reusable and in which the action of the substance is ancillary to that of the device	condoms with spermicidal
	Devices used in contraception, invasive, implantable, or long-term use	intrauterine device, which does not release progestogens

Portuguese national safety alerts on medical devices in 2017

Figure 1. Type of public safety alerts on medical devices in Portugal (2017)



DISCUSSION

Brief Analyzes

Overall, there were few alerts on medical devices released and published in the official website of the Portuguese medicine agency during 2017 (i.e. only thirty-two alerts). This limited number of alerts may be associated with a limited number of marketed devices in Portugal, considering that the Portuguese economy have recently faced restrictions, but further studies are required to support this assumption. More than a half of alerts were on medical devices without identified records of commercialization in the Portuguese territory, since EU is an open marketed where goods, persons and services can circulate for free (EU, 2019).

Notified adverse incidents with medical devices are evaluated and, if relevant (e.g. severe incidents), they are circulated in the form of a NCAR. Only 0.3% of all NCAR were reported in Portugal. This finding is consistent with the limited number of alerts on medical devices identified at a national level (EU statistics, 2017). In 2017, the three most prevalent occurrences from the 981 (100%) NCAR exchanged between 17 EU countries were: 363 (37.0%), 190 (19.3%), and 115 (11.7%) in Germany, United Kingdom, and France, respectively (EU statistics, 2017). Thus, it seems that medical devices are safely used within the Portuguese market, although the extent of information available in the safety alerts of medical devices seems to be limited. For instance, the possible patients' safety consequences were not specifically described in the medical device alerts.

The types of instructions/actions taken by the competent authorities were variable depending on the type of alert. For instance, as expected counterfeit devices were withdrawn from the market. New actions/safety measures may be required by competent authorities to avoid the commercialization of these counterfeit devices even in EU, which is an extremely regulated market. Marketing surveillance of medical devices is fundamental to assure patient safety. In this sense, alerts to not purchase or use medical devices without identified records of commercialization are relevant, since within the European economic area, goods and people are free to circulate/move.

Solutions and Recommendations

The traceability of medical devices is a relevant issue, since people and goods move across borders (European Union, 2019). Therefore, it could be advisable that retail sellers of medical devices, such as ambulatory or hospital pharmacies, could collect patient data upon acquisition of the medical device and, if necessary, be responsible for warning the patient in case of new information or safety concerns related to the dispensed medical device. More public information on how to report adverse events in medical devices and how to consult safety alerts for medical devices should be created and disseminated. For instance, a national campaign on the safety of medical devices could be an important instrument to highlight the need to report potential adverse reactions and to raise awareness of the safety issues involving these devices. Furthermore, in order to avoid the introduction of counterfeit medical devices in the market, more straightforward supervision may be required.

Future Research Directions

In the future, all medical devices (or their packages) may be associated with an electronic mechanism to allow traceability and the recognition of a medical device with related safety issues, as a way of emitting alerts. The development of economic, safe, and efficient medical devices is relevant to assure patients access to these health products. New classifications and regulations on medical devices may be required if new classes of medical devices appear. Citizens and patients may be questioned on the usability and intelligibility of medical devices. In EU, including in Portugal the impact of the application of the new Directive 2017/745 should be monitored, to allow the analysis of a possible better risk-benefit profile of the new approved medical devices through the application of the new Directive. Also, comparative studies of clinical safety on approved medical devices (e.g. between EUA e UE) should be carried out to support the robustness of the methodology used in high-risk trials to approve medical devices.

Finally, it will have a major impact the development of a longitudinal study comparing the number of alerts reported in Portugal with the data in another EU countries and the total number of alerts reported in the EU for medical devices.

CONCLUSION

In 2017, a small number of safety alerts on medical devices were identified in Portugal. A significant part of these alerts were related to devices not market in Portugal, which may have contributed to this low number of alerts. The types of instructions/actions taken by the competent authorities depended on the level of risk associated with the use of a certain device. As expected, the most serious events (or potentially serious events) were associated with withdrawals of devices from the market, such as in the case of falsified devices. Although the number of identified counterfeit devices was very limited, this situation represents a potential public health risk/concern.

ACKNOWLEDGMENT

This work is supported by national funds from FCT – Fundação para a Ciência e a Tecnologia, I.P., under the project UID/DTP/04567/2016.

REFERENCES

Cochrane Library. (2019). *About the Cochrane Library*. Retrieved from: <https://www.cochranelibrary.com/about/about-cochrane-library>

Council Directive 90/385/EEC active implantable medical devices. *Official Journal of the European Communities*, 189, 17–36. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31990L0385&from=EN>

Council Directive 93/42/EEC concerning medical devices. *Official Journal of the European Communities*, 169, 1–43. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31993L0042&from=PT>

Council Directive 98/79/EC on in vitro diagnostic medical devices. *Official Journal of the European Communities*, 331, 1–37. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31998L0079&from=en>

Davis, S., Gilbertson, E., & Goodall, S. (2011). *EU Medical Device Approval Safety Assessment: A comparative analysis of medical device recalls 2005-2009*. The Boston Consulting Group. Retrieved from: http://www.medtecheurope.org/sites/default/files/resource_items/files/01012011_BCG_EU%20medical%20device%20approval%20safety%20assessment_Backgrounder.pdf

Decree- Law No. 145/2016 on medical devices. *Decreto- Lei n.º 145/2016*. Retrieved from: <https://dre.pt/web/guest/pesquisa/-/search/494558/details/maximized>

European Commission. (2001). *EUDAMED: European Database on Medical Devices*. Retrieved from: <https://ec.europa.eu/idabc/en/document/2256/5637.html>

European Commission. (2018). *Medical devices*. Retrieved from: https://ec.europa.eu/growth/sectors/medical-devices_pt

European Parliament and of the Council. (2017). *Directive 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices*. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745>

European Union. (2017). *2017 NCARs - National Competent Authority Report: EU statistics*. Author.

European Union. (2019). *One market without borders*. Retrieved from: https://europa.eu/european-union/topics/single-market_en

Food and Drug Administration. (2018a). Federal Food, Drug, and Cosmetic Act [As Amended Through P.L. 115–271, Enacted October 24, 2018]. Retrieved from: <https://legcounsel.house.gov/Comps/Federal%20Food,%20Drug,%20And%20Cosmetic%20Act.pdf>

Food and Drug Administration. (2018b). *Is The Product A Medical Device?* Retrieved from: <https://www.fda.gov/medicaldevices/deviceregulationandguidance/overview/classifyyourdevice/ucm051512.htm>

Gail, A., & Norman, V. (2016). Drugs and Devices: Comparison of European and U.S. Approval Processes. *JACC. Basic to Translational Science*, 1(5), 399–412. doi:10.1016/j.jacbts.2016.06.003 PMID:30167527

Global Harmonization Task Force. (2012). *Definition of the Terms 'Medical Device' and 'In Vitro Diagnostic (IVD) Medical Device'*, *GHTF/SG1/N071*. Retrieved from: <http://www.imdrf.org/docs/ghtf/final/sg1/technical-docs/ghtf-sg1-n071-2012-definition-of-terms-120516.pdf#search=>

Gruenwald, J. (2017). *New Medical Device Regulation Update*. Retrieved from: https://www.nutraceuticalsworld.com/contents/view_columns/2017-06-02/new-medical-device-regulation-update/

Hwang, T. J., Sokolov, E., Franklin, J. M., & Kesselheim, A. S. (2016). Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: Cohort study. *BMJ (Clinical Research Ed.)*, 353, i3323. doi:10.1136/bmj.i3323 PMID:27352914

INFARMED. I.P. (2016). *Medical devives in Pharmacy*. Retrieved from: http://www.INFARMED.pt/web/INFARMED/entidades/dispositivos-medicos/aquisicao-e-utilizacao/dispositivos_medicos_farmacia

International Trade Administration. (2016). *2016 Top Markets Report Medical Devices*. Retrieved at: https://www.trade.gov/topmarkets/pdf/Medical_Devices_Top_Markets_Report.pdf

Jarow, J. P., & Baxley, J. H. (2015). Medical devices: US medical device regulation. *Urologic Oncology*, 33(3), 128–132. doi:10.1016/j.urolonc.2014.10.004 PMID:25458071

Kramer, D. B., Tan, Y. T., Sato, C., & Kesselheim, A. S. (2013). Postmarket Surveillance of Medical Devices: A Comparison of Strategies in the US, EU, Japan, and China. *PLoS Medicine*, 10(9), e1001519. doi:10.1371/journal.pmed.1001519 PMID:24223035

Lucintel. (2016). *Growth Opportunities in the Global Medical Device Market*. Retrieved at: <https://www.lucintel.com/medical-device-market-2016-2021.aspx>

Olberg, B., Fuchs, S., Panteli, D., Perleth, M., & Busse, R. (2017). Scientific Evidence in Health Technology Assessment Reports: An In-Depth Analysis of European Assessments on High-Risk Medical Devices. *Value in Health*, 20(10), 1420–1426. doi:10.1016/j.jval.2017.05.011 PMID:29241902

Peña, C., Bowsher, K., Costello, A., De Luca, R., Doll, S., Li, K., Schroeder, M., & Stevens, T. (2007). An overview of FDA medical device regulation as it relates to deep brain stimulation devices. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(3), 421–424. doi:10.1109/TN-SRE.2007.903973 PMID:17894274

WHO. (2018). *Definitions: medical device (brief)*. Retrieved from: https://www.who.int/medical_devices/definitions/en/

World Bank Income Group. (2016). *Medical devices regulatory systems at country level: Portugal*. Retrieved from: https://www.who.int/medical_devices/countries/regulations/prt.pdf

ADDITIONAL READING

Heneghan, C., Thompson, M., Billingsley, M., & Cohen, D. (2011). Medical-device recalls in the UK and the device-regulation process: Retrospective review of safety notices and alerts. *BMJ Open*, 1(1), e000155. doi:10.1136/bmjopen-2011-000155 PMID:22021778

Hidefjäll, P., & Backheden, M. (2017). Making healthtechnologyassessmentmore dynamic – Temporal trend analysis to capture performance trajectories. *Health PolicyandTechnology*, 6, 328:338.

Markiewicz, K., Til, J. V., & IJzerman, M. (2017). Early Assessment of Medical Devices in Development for Company Decision Making: An Exploration of Best Practices. *Journal of Commercial Biotechnology*, 23(2), 15–30. doi:10.5912/jcb780

O'Dwyer, C., & Cormican, K. (2017). Regulation – Do or Die: An analysis of Factors Critical to New Product Development in a Regulatory Context. *J. Technol. Manag. Innov.*, 12(1), 26:38.

Privitera, M.B., Evans, M., & Southee, D. (2017). Human factors in the design of medical devices e Approaches to meeting international standards in the European Union and USA. *Applied Ergonomics*, 59, 251e263.

Quinn, P. (2017). The EU commission's risky choice for a non-risk based strategy on assessment of medical devices. *Computer Law & Security Review*, 33, 361:370.

Skerritt, J. (2017). Medical device regulatory reforms 2017: a period of significant change. *Australasian BioTechnology*, 27(1), 38:42.

Vincent, C.J., Li, Y., & Blandford, A. (2014). Integration of human factors and ergonomics during medical device design and development: It's all about communication. *Applied Ergonomics*, 45, 413e419.

KEY TERMS AND DEFINITIONS

Competent Authority: A public authority with specific competences in the commercialization medical devices.

Class I-III: The classes/classifications of medical devices in accordance to Annex of the IX, Council Directive 93/42/EEC.

Directive: A regulation/law in European Union.

EUDAMED: A European database comprising commercial and safety information on medical devices.

INFARMED, I.P.: The Portuguese medicine agency.

Medical Device: A device for medical use, which may not be classified as a medicine.

Notified Bodies: A private entity responsible for certifying medical devices in accordance to law.

Note: CBIOS - Universidade Lusófona's Research Center for Biosciences and Health Technologies, Campo Grande, 376, 1749-024, Lisboa, Portugal.

ENDNOTES

¹ Cochrane Library is a collection of databases that contain evidence information to support health-care-decision making and other (Cochrane Library, 2019).

<https://www.cochranelibrary.com/search> (accessed in 2019-03-10)

² <http://www.INFARMED.pt/web/INFARMED/alertas/dispositivos-medicos>.

Medical Social Networks, Epidemiology and Health Systems

15

Patrícia C. T. Gonçalves <https://orcid.org/0000-0003-4253-7474>*INESC TEC, University of Porto, Portugal***Ana S. Moura***LAQV-REQUIMTE, University of Porto, Portugal***M. Natália D. S. Cordeiro** <https://orcid.org/0000-0003-3375-8670>*LAQV-REQUIMTE, University of Porto, Portugal***Pedro Campos***INESC TEC, University of Porto, Portugal*

INTRODUCTION

Pandemic crisis, location of the primary case and individual medical data are all major concerns and features in the global health environment. Pandemic crisis, *i.e.*, an epidemiological outbreak that is simultaneously on several continents, with all the social, economic and individual mayhem it provokes, links tightly with the location of the primary case, which is the individual who introduces the disease into the population, and whose clinical data, *i.e.*, the genetic and personal health related data, may prove invaluable to assess the global health policies and countermeasures on the crisis.

These three aspects are among some of the Health and Social problems that the emerging medical social networks can positively address. Medical networks, which work as an interface between the patient medical data and his/hers geographical and/or social connections, as well as between the patient's individual needs and the attending medical doctor, can allow feasible and fast visualization/information systems. Furthermore, they are effective and emerging research tools in studying health policies and detecting how to improve them.

In the recent five years, there has been an increased research interest in the development and application of informal medical social networks, as well as data concerning the real advantages of their use on daily situations and Health Policies Implementation. In 2018, an ideal model for a medical social network and data visualization information system has been proposed (Gonçalves, Moura, Cordeiro & Campos, 2018), which considers both the individual and the socio-economic needs and rights, and it is likely that the first prototypes can be presented in the near future.

This chapter reviews the evolution and *status quo* of prospective medical social networks, and discusses the ideal model and its future venues and interaction with ethics in the areas of Economics, Health Policies and Human Rights.

DOI: 10.4018/978-1-7998-3479-3.ch126

BACKGROUND

The advent of the concept ‘medical social network’ came in 2007, when Barabási published a work entitled *Medicine Network: From Obesity to the “Diseasome”* (Barabási, 2007). Though Barabási used the term ‘medicine network’ instead of ‘medical social network’, he defended that since networks can be found on every particulars and features of human health, as knowledge regarding its effects on the patients’ biomedical aspects increased, a new field of expertise on medical practice and research would appear. These medical social networks were constructed and studied through graph theory visualization, namely as nodes and edges (Newman, 2010).

A node, within a social network, is defined as a distinct entity, which can be single, like a person or a molecule, or collective, like a country or a company. An edge, in the same context, represents the connection between two different nodes. The connections are multi-natured, *i.e.*, they can be friendship or working connections, for example, if one thinks of human individuals; but they can also be of chemical definition, for example, as the bond type between intramolecular atoms. Exemplifying, when restricting to the medical context, the nodes may represent people, coupled by their biological or psychological factors, such as obesity (Christakis & Fowler, 2009) or depression (Rosenquist, Fowler, & Christakis, 2011), but they can also represent diseases themselves, like in comorbidity studies, *i.e.*, the causal interaction between diseases (Folino, Pizzuti & Ventura, 2010). In the same context, the edges are the type of connections between the people, *i.e.*, genetic/familiar or geographical, or different causal levels of comorbidity, if the nodes are diseases. These simple examples do not exempt the graph model to have added complexity, since the nodes can also comprise the information of gender, for instance, apart from the genetic or otherwise bio, social, or psychological factors. Figure 1 illustrates in a simplified manner the relation between real interactions and corresponding graphs.

Figure 1. Example of a multi-layer social network representing real life connections between people. (Image obtained with software MuxViz(De Domenico, Porter & Arenas, 2015).)

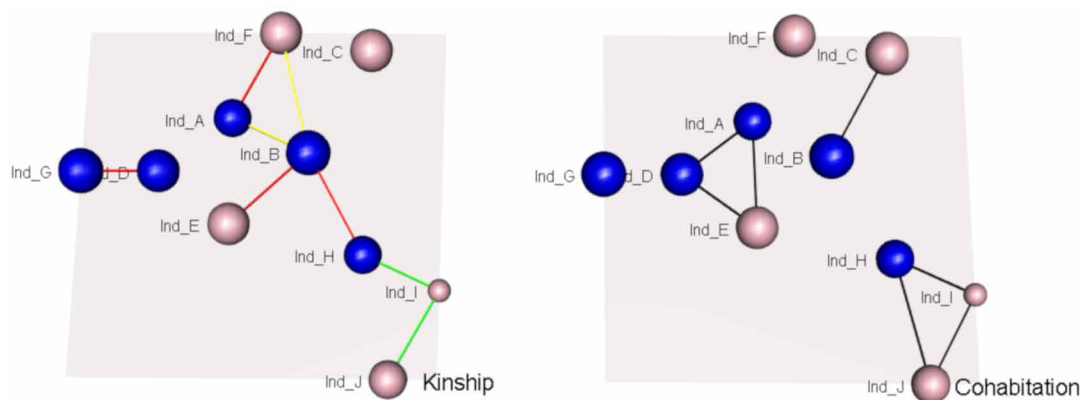


Figure 1 represents a social network with two layers. The left, or first, layer represents the kinship between the nodes, which are individuals, and the right, or second, layer represents cohabitation connections. The pink nodes represent female individuals and the blue nodes represent males. The size of the nodes is proportional to the age of the individual, and in the left layer, the edge colour represents different kinds of family connections. Yellow stands for grandparent/grandchild connection, green stands for

parent/offspring ties, and, in red, one has sibling relationships. As the right layer stands for one type of connection only, *i.e.*, cohabitation, all edges have the same colour. This is a type of social network that adapted for medical purposes could allow for contagion detection in several types of diseases, identifying those from direct contact or genetic/hereditary.

Indeed, almost fifty years ago, the importance of meeting health information with global population knowledge had already been signalled as fundamental for developing public health policies and sound implementation near the populations (Wennberg & Gittelsohn, 1973). Decades later, when the Internet was already becoming established as a worldwide means of information for the global population, with the increasing popularity of instant messaging programs, such as the extinct AOL Instant Messaging (AIM), again this idea was reinforced, by researchers, stating that such a type of synergetic information system could aid both national and regional health policies, by monitoring efficient allocation of medical resources, either human or logistic, and adapting those medical resources to the citizens' different idiosyncrasies, from health status to socioeconomic risk, and access to the several medical resources (Roos *et al.*, 1999).

The past fifteen years witnessed the emergence of several models of medical social networks for applications ranging from analysis of public health policies to locating patient zeros on epidemic crisis, and even studying contagion on conditions such as depression or obesity. The present state of application and use of these models, as well as their theoretic and practical principles, are discussed on the following sections, per type of application, *i.e.* if they are applied in public health policies monitoring or analysis of disease contagion, for example.

FOCUS OF THE ARTICLE

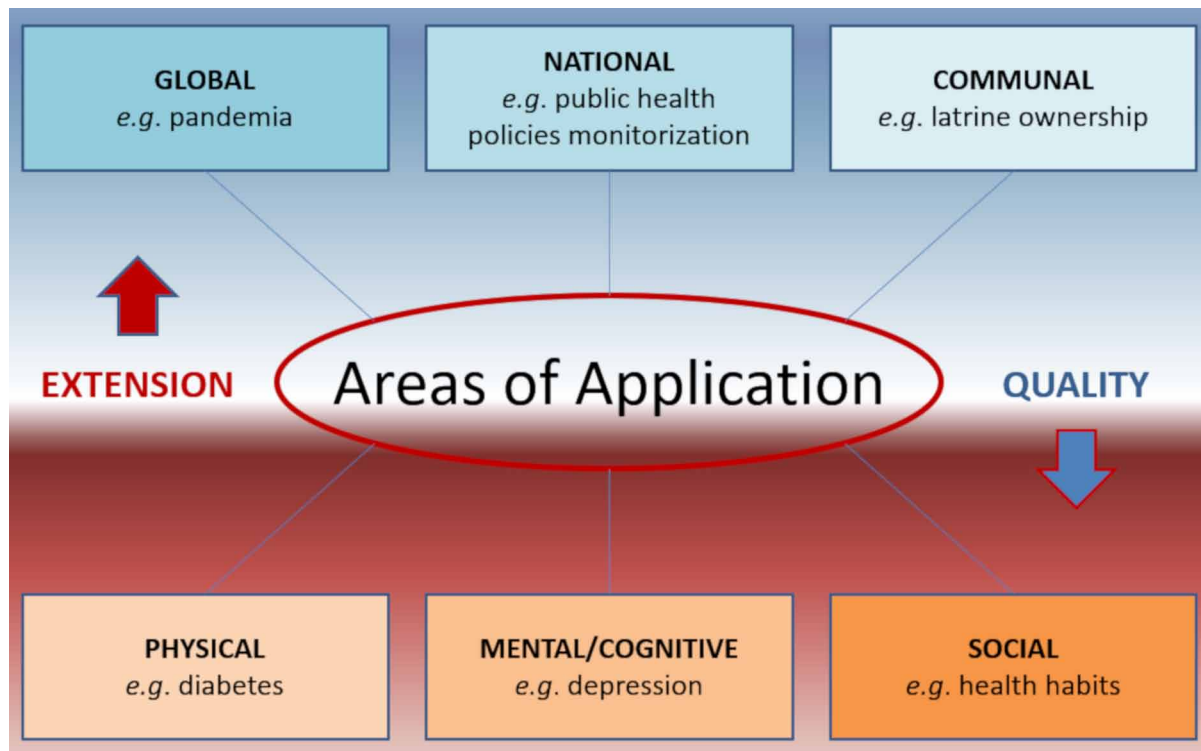
Getting insight into the many areas of application and scope of utility of medical networks can be made more feasible by discriminating such areas into two major groups, one where the context is the human extension of the population composing the medical network, *i.e.*, the order of magnitude of the population, and the other context is the type of human quality being analysed, *i.e.*, the dependent variable, either quantitative or qualitative, under scrutiny. In the first case, one can broadly consider three main population magnitudes: global, national, and community wide. On the second context, one can also consider three main quality types: physical, mental/cognitive and social. All of these, whose connections are illustrated in Figure 2, are discussed in the following subsections in the form of relevant case studies.

With this classification as guidance, the selection of the literature to be discussed followed. Several criteria were applied on this selection: (a) only one publication *per* the above defined categories; (b) the chosen publication should be a clear example of the category, with the purpose of practical application and, ideally, with proven results; (c) in case of an equal situation for two publications regarding (a) and (b), the publication with the most relevance for the development and implementation of social medical networks, seen by citation and/or practical implementation, was chosen; and (d) the final criteria would be the most recent publication, if similar status was met by two publications within (a), (b) and (c) context. As such, six case studies, one *per* category, were selected and are discussed in the following subsections.

Medical Social Networks: Global Order of Magnitude

The use of social networks for medical purposes can also depart from plain social networks *per se*. In fact, online apps for social interaction, such as Facebook, Instagram or Twitter, represent an excellent

Figure 2. Areas of application of Medical Social Networks, according to the order of magnitude of the populations being addressed, and according to the type of health feature in study.



sample of global population as they have, respectively, 2.27 billion (Statista, 2018), 1 billion (Omnicores, 2018), and over 300 million (The Balance Careers, 2018) monthly active users. Given the various type of personal data that can be collected from such social networks it was inevitable that medical applications would resource to it. A recent and interesting research explored the eventual relation between the use of social media and personal well-being (Shakya & Christakis, 2017). The work departed from the premise that non-virtual interaction enhances personal well-being and explored if such a conclusion could also be reached for virtual active relationships. Within four domains of well-being, namely body mass index, self-reported life satisfaction, self-reported mental and physical health, the objective public data from Facebook users was cross referenced with three waves of self-reported measured social data from non-virtual social networks.

The researchers concluded that, within some significance variances, a rather robust pattern was identified. Real-life social networks are positively associated with overall well-being, while Facebook use is negatively associated. Nevertheless, they also point out that the negative associations with Facebook use were stronger than the positive associations of real-life interactions with overall well-being. They also asserted that the positive interaction with other Facebook users' content, *i.e.*, 'liking' posts or links of other users, was consistently related with a more compromised overall well-being, while the quantification of Facebook updates, *i.e.*, the number of status updates of each user, was associated with diminished mental health. Other aspects, such as greater number of Facebook friends, *i.e.*, the user's Facebook network, though evidencing positive association with well-being in cross-sectional models, lost its significance when within a longitudinal model. A hypothesis laid by the authors was also the

mutually exclusive nature of offline versus online relationships as the negative associations of Facebook and positive correlations of real-world interactions were of comparable extension.

The use of social networks data in medical social networks as an effective monitor of several aspects of well-being was demonstrated (Shakya & Christakis, 2017). However, it should be noted the following: other than public data Facebook users might provide, the interaction data is subject to permission and that affected the longitudinal data size, and the study was unable to identify the mechanism through which Facebook use causes reduced well-being. Though the second can be, and is being, explored in new studies, the first is of extreme importance. Data protection is becoming a more urgent matter, as scandals associated with the misuse of social networks ravel one after the other. Though the right for privacy is undisputed, the interface between such right and the medical access to private data collected through social network apps must be thoroughly defined and with optimal consequences for both parts. Such aspects will be addressed on the section of *Future Perspectives*.

Medical Social Networks: National Order of Magnitude

The potential of developing a health integrated system of information for medical practitioners and Public Health decision-makers alike had already been duly right before the turn of the century (Roos et al., 1999). This work, though not formally introducing data from social network apps into a medical social network, since at that time the online social media was still an Internet toddler, it proposes a synergetic network model that illustrates how present-day data can be integrated in a medical social network.

The authors departed from the traditional overall assumptions that coordinated the allocation of hospital services in national and regional scenarios, *i.e.*, population growth associated with health policy decisions of increasing or allocating new hospital services in the growth region or increased demand on volume of use. They also took into consideration that the care assessment focused the individual results of treated citizens or the rating of hospital institutions, and pointed out this was rooted on the conviction that across the overall populations, there was a similar need for specific health care.

However, many studies seemed to challenge these assumptions and in response they elaborated a model that proposed a comprehensive population health information system whose administrative database cross-referenced the following areas: location of allocated care (medical clinics, nursing home, hospital institution, citizen home), overall population statistics (including census), health-related services data (as from pharmaceutical services, for example), and physicians data. The model linked four sets of population information (population health, socioeconomic risks characteristics, health care expenditures, supply of health care resources) and would organize the information per geographical area according to these indicators. Another important feature of the model was how both policymakers and the general public would be able to discriminate factors within the health issues as independent social levers outside formal health care systems could be identified (Roos et al., 1999).

The authors conclude by stating that the capabilities of such an information system only now are being explored, and that the specific system they present creates a dynamic interface between academics, policymakers and the general public. In fact, at present, such benefits are becoming a reality with the inclusion of these types of data in prototypes of medical social networks.

Medical Social Network: Communal Order of Magnitude

The third case study within the extension variable regarding the applications of medical social networks, used an algorithmic method applied to such networks in order to evaluate how several Indian communities

changing a specific health habit, which was the ownership of a latrine, would not only affect both the community and village level of latrine ownership but also how the network cohesion affected individually the villagers in such matter (Shakya, Christakis, & Fowler, 2014).

The researchers explored both descriptive, *i.e.*, communal regular behaviour adopted without other's expectations, and injunctive, *i.e.*, communal standards enforced by social sanctions, norms within the context of individual latrine ownership *versus* community health habits change. They considered three social levels, friendship, social network community and overall village social network. It became clear that there was a strong association between owning an individual latrine and changes in all three levels, albeit the strongest effect was found at the social network community level. In fact, the higher the level of community network cohesion, the lower the probability of individual latrine ownership, though this only applied to the cases where the number of individual latrine ownership were less significant.

The authors concluded that health and international development interventions in such context should be determined not by geography but using social network knowledge, breaking traditional statistical data approach to a more humane and intelligible health policy.

Medical Social Networks: Physical Variables

This entry marks the first case study of medical social networks within the context of human quality types. Human health presents several aspects, and the most obvious since the early days of medicine was the physical aspect, *i.e.*, the overall aspects of the bodily parts and functions of a human being. This research studied how a prevention of a physical infirmity as a cardiovascular event and the adoption of preventive measures by social interactions could be associated (Strully et al., 2012). The authors considered the aspirin intake as the preventive measure and used a longitudinal logistic regression model to explore such scenario in three data waves extracted from the Framingham Heart Study. The study also included several individual aspects, such as gender or type of social connection, and approached the results through traditional statistical functions, displayed in tables.

The conclusions had two main aspects. Regarding the premise of the study, the sharing of medical doctor was a common feature when the aspirin intake habits were changed. The second aspect was the conclusion by the authors that such a traditional display of results made difficult to locate the individual aspects that sustained such changes. For example, and taking into consideration the identification of the common feature, *i.e.*, the sharing of doctor in individuals changing habits regarding aspirin intake as a preventive measure for cardio vascular events, it was not possible to determine if the doctor had actively influenced the decision. The authors themselves pointed out that the non-inclusion of a social network model could be the limiting factor in identifying such aspect, as well as others (Strully et al., 2012).

Medical Social Networks: Mental/Cognitive Variables

Depression, as other mental infirmities, had empirically been thought to be associated with social factors. However, the inclusion of social network models to conclude regarding health care in such context has only been studied in recent years, as in this case study, which had a sample of a more than ten thousand people social network (Rosenquist et al., 2011). Furthermore, the sample was assessed several times over a timespan of more than three decades, and longitudinal statistical models were used to probe how individual depressive symptoms were related with similar symptomatic levels of their social network. The personal connections would vary from the genetic liaison, as in siblings, for example, to intimate, as in spouses, to geographical, as in neighbours.

Correlation was found between depressive symptoms and intimacy and geographical connections, *i.e.*, proximity to the home of the depressed individual would affect the depressive levels of the others. Gender was also a variable to be considered as fulcrum, since being cumulatively a female within the proximity would seem more susceptible to be influenced. The results were robust through several simulations of the network and other methodologies and reveal the contagion features of depressive states (Rosenquist et al., 2011).

The authors stated that the inclusion of social network models for medical data would allow to not only better understand the impact of interpersonal relationships on mental conditions but also how to develop future health policies that could locate and prevent the conditions for contagion.

Medical Social Networks: Social Variables

The final of the selected case studies regards the social aspects of human health and human health policies. In this work, the question in wanting of answer was if health policies targeting the change of health habits, namely in nutrition, that focused in the public example of influential individuals were efficient. In order to obtain an answer to this question and being able to develop health policies that maximized the level of population behaviour change, the authors recruited individuals from thirty-two Honduran villages and discriminated the villages on network size, socioeconomic states, and baseline rates of water purification for two intended public health inversions, one regarding the use of chlorine in water purification, the other, regarding multivitamins intake for micronutrients deficiencies (Kim et al., 2015).

It was concluded that change in health habits was far more efficient through close personal connections rather than through the example of influential individuals. The intervention diffusion could be maximized by analysing the properties of the human social network. The authors clearly indicated that the adoption of social network models to health-related data could be a more sustainable and efficient manner to develop public health policies and interventions.

SOLUTIONS AND RECOMMENDATIONS

On a recent publication, an ideal model for the principles and day-to-day application of a medical social network was suggested (Gonçalves *et al.*, 2018). Through a story-type scenario, the model proposes the ideal interaction between a physician and a hypothetical patient, Mr. Silva. Mr Silva has a pre-scheduled appointment with his personal doctor, being it on a private or public medical institution is irrelevant for the model. Due to this pre-scheduling, the doctor can access all the data the patient has previously authorized to be available for the medical professionals dealing with his case. This data includes, for example, standard clinical data such as doctor appointments records, blood tests results, or X-ray images, but also social network data.

The specifics of social network data available to the doctor are: (a) relevant medical data for Mr. Silva's health status of kinship, *i.e.*, of the first degree relatives such as parents, siblings, or offspring, and second degree relatives, such as uncles, aunts, grandparents or full cousins; and (b) relevant public alert on medical networks, either national or local, affecting Mr. Silva's (i) working relations and (ii) neighbourhood interactions. This would allow to cross-reference the patient's personal medical records with family medical history and social interactions. As such, the physician is not only up to date with much (though not all) of the needed information to support the diagnosis of Mr. Silva, even prior to his entering in the room, but is also in a position to act as a public health agent for disease prevention on

a local and national scale. Further, and reinforcing the authorization Mr. Silva must give for the availability of the above-mentioned data, the right to privacy is guaranteed as the doctor only has access to the type of relationships and respective clinical data but never to the identity of the individuals of Mr. Silva's social networks (Gonçalves *et al.*, 2018). On an eventual public health crisis, the identity of the connections could be revealed, after external authorization, taking in consideration all present legal and human rights regarding privacy.

The public interest and purpose of the ideal model would be sustained by the continuous synergetic work of professionals from areas such as data management and analysis, information technology (IT) analysis and health practitioners which would cross-reference an array of medical social networks with the detection of potential health threats. These threats could be either of sporadic situations, such as epidemics, or ongoing, such as depression (Rosenquist *et al.*, 2011).

An interesting recommendation for upgrading this model would be to have the type of data regarding Mr. Silva's kinship elaborated for the adjacent geographical connections, established by a pre-defined radius for circular areas around his home and/or workplace. Further, the model should be accompanied by the implementation of mandatory instructions for medical professionals in order to be understood not only how it works but also what data is still needed to extract from the patient, *i.e.*, it should be clear that the ideal model is not a diagnosis system but a supporting system for the diagnosis.

The aspect of privacy within the social medical networks can be considered a research field on its own, and more so with the new pivotal role of artificial intelligence (AI) on IT. AI implies that algorithms, *i.e.*, the specific process on a given computer operation, can be developed to emulate human performance on tasks such as decision-making, and if one adds that to the emergent AI field of machine learning, which allows the computer systems to automatically improve through the identification of problems and auto-correction, the concerns regarding the misuse of personal and medical data available on medical social networks are real and need to be seriously addressed.

As such, and collecting from public and governmental documents, as the 'Guidelines on Artificial Intelligence and Data Protection' of the Council of Europe, one can identify the following main aspects that need to be incorporated in the ideal model (Convention 108, 2019): (a) according to the principles of Convention 108+, the ideal software model should incorporate privacy-by-design and by default, while taking into account the proportionality of data processing and the purpose specification, and policy makers should impose on AI developers, manufacturers, and service providers the mandatory need to assess the impact of data processing of the developed software on fundamental freedoms and rights, namely the right to privacy, before the product implementation (Council of Europe, 2018); (b) the risk assessment of these softwares should adopt a multiview of possible outcomes of the data processing, in order to be aligned with human rights, fundamental freedoms and the functioning of democracies and social and ethical values – this aspect is particularly important in the development of an ideal model of a medical social network, as the application presupposes scenarios where the greater good versus individual freedom may be present, as in the case of epidemics (Council of Europe, 2017); (c) as all human interactions provide and are recipient of meaningful transitions, from a cognitive and emotional perspective, these AI medical applications should allow the data subjects, *i.e.*, the human subjects, to interact with the data processing and its effects with meaning, that is, the process should not be developed in such a manner that it increases the hypothetical software illiteracy of the subjects; and (d) the subject's right to intervene on the AI decisions should be preserved, and the AI developers, manufacturers, and service providers should be responsible and accounted for adopting an algorithm vigilance which assures the applications respect the data protection principles and human rights.

A final recommendation, and taking into consideration that the contextual nature of this aspect is multidisciplinary, it would be interesting to propose the development of a ‘Good Practices Handbook’ for the elaboration of such medical social networks, which would comprise the insights of three different parties: academics and researchers, policy-makers, and the civil society, each with equal representation on the adoption of the measures. The discussion of the measures and problems should be two-folded: through online forum debates and physical gatherings with at least one element of each party.

FUTURE RESEARCH DIRECTIONS

The recent emergence of published research in the area of social networks and their application on Health has indicated both the failures of the software paths and goals for the prospective systems. In fact, it has been noted that software companies’ efforts over the last decade on the Health area have not eliminated medical errors as medical information is scattered instead of transversally integrated (Carroll, 2016). Indeed, as medicine is increasingly taking the path from reactive partner to health incidents, either individual or collective, to become a proactive agent, with a one-to-one preventive interaction between medical practitioners and citizens, the need to access integrated healthcare information is a cornerstone (Carroll, 2016). Further, the value of such synergy between the several parts in the health area, either as providers, such as hospitals, or recipients, such as payers, and even as both in different contexts, such as public health departments, has been positively assessed in the last two decades (Walker et al., 2005).

One can point out also that with the recent laws regarding privacy, such as European Union’s General Data Protection Regulation, and reinforced focus on encrypted data for security reasons of the rights of the common citizen, that the intersection of interests between medical social networks, medical information, public health policies and data security will become a field of research in the next two years, and its results taken into account in public health protocols.

Further, from a sociological perspective, social medical networks can simultaneously harvest from and deliver new seeds to the understanding of social capital and its applications, a field focusing on networks of people in a given society, as the concepts at its roots are akin to the social values and interrelationships between the individuals of a community. In fact, an indication that this emerging field could find an interesting spot in this research area is the number of studies carried out on health and social capital, which grew twenty times between 1995 and 2006 (Kawachi, Subramanian & Kim, 2008).

CONCLUSION

Medical social networks are simultaneously the frontier and avenue of effective health care systems. They allow a fast and feasible means to assess health care quality and monitor the three human components of the scenario: decision-makers, general population, and health professionals. However, and though a plethora of works presenting the advantages of cross-referencing social network data with online medical (and) social network data have been published in recent years, some questions still need to be addressed in order that the solution does not become a Kafkaesque nightmarish reality, namely the protection of the individual right to privacy. Within or outside the traditional frameworks of data protection, new proposals need to be made, while the system smoothly advances in scattered areas.

REFERENCES

- Barabási, A.-L. (2007). Network Medicine— From Obesity to the “Diseasome”. *The New England Journal of Medicine*, 357(4), 404–407. doi:10.1056/NEJMe078114 PubMed
- Carroll, N. (2016). Key Success Factors for Smart and Connected Health Software Solutions. *Computer*, 49(11), 22–28. doi:10.1109/MC.2016.340
- Christakis, N. A., & Fowler, J. H. (2009). Social Network Visualization in Epidemiology. *Norsk Epidemiologi*, 19(1), 5–16. doi:10.5324/nje.v19i1.6 PubMed
- Convention 108. (2019). Guidelines on Artificial Intelligence and Data Protection. Directorate General of Human Rights and Rule of Law, Council of Europe.
- Council of Europe. (2017). *Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in a World of Big Data*. Directorate General of Human Rights and Rule of Law. Council of Europe.
- Council of Europe. (2018). Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Council of Europe Treaty Series No. 223.
- De Domenico, M., Porter, M. A., & Arenas, A. (2015). MuxViz: A Tool for Multilayer Analysis and Visualization of Networks. *Journal of Complex Networks*, 3(2), 159–176. doi:10.1093/comnet/cnu038
- Folino, F., Pizzuti, C., & Ventura, M. (2010). A Comorbidity Network Approach to Predict Disease Risk. Springer., doi:10.1007/978-3-642-15020-3_10
- Gonçalves, P. C. T., Moura, A. S., Cordeiro, M. N. D. S., & Campos, P. (2018). Mr. Silva and Patient Zero: A Medical Social Network and Data Visualization Information System. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation* (pp. 111–117). Lecture Notes in Computer Science. Springer International Publishing., doi:10.1007/978-3-030-01045-4_13
- Kawachi, I., Subramanian, S. V., & Kim, D. (2008). Social Capital and Health. In *Social Capital and Health* (pp. 1–26). Springer New York. doi:10.1007/978-0-387-71311-3_1
- Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O’Malley, A. J., Fowler, J. H., & Christakis, N. A. (2015). Social Network Targeting to Maximise Population Behaviour Change: A Cluster Randomised Controlled Trial. *Lancet*, 386(9989), 145–153. doi:10.1016/S0140-6736(15)60095-2 PubMed
- Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press., doi:10.1093/acprof:oso/9780199206650.001.0001.
- Omnicore. (2018). Instagram by the Numbers: Stats, Demographics & Fun Facts. Retrieved 30 November 2018, from <https://www.omnicoreagency.com/instagram-statistics/>
- Roos, N. P., Black, C., Roos, L. L., Frohlich, N., DeCoster, C., Mustard, C., Brownell, M. D., Shanahan, M., Fergusson, P., Toll, F., Carriere, K. C., Burchill, C., Fransoo, R., MacWilliam, L., Bogdanovic, B., & Friesen, D. (1999). Managing Health Services: How the Population Health Information System (POPULIS) Works for Policymakers. *Medical Care*, 37(6, Suppl), JS27–JS41. doi:10.1097/00005650-199906001-00007 PubMed

Rosenquist, J. N., Fowler, J. H., & Christakis, N. A. (2011). Social Network Determinants of Depression. *Molecular Psychiatry*, 16(3), 273–281. doi:10.1038/mp.2010.13 PubMed

Shakya, H. B., & Christakis, N. A. (2017). Association of Facebook Use With Compromised Well-Being: A Longitudinal Study. *American Journal of Epidemiology*, 185(3), 203–211. doi:10.1093/aje/kww189 PubMed

Shakya, H. B., Christakis, N. A., & Fowler, J. H. (2014). Association Between Social Network Communities and Health Behavior: An Observational Sociocentric Network Study of Latrine Ownership in Rural India. *American Journal of Public Health*, 104(5), 930–937. doi:10.2105/AJPH.2013.301811 PubMed

Statista. (2018). Number of Monthly Active Facebook Users Worldwide as of 3rd Quarter 2018. Retrieved 30 November 2018, from <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>

Strully, K. W., Fowler, J. H., Murabito, J. M., Benjamin, E. J., Levy, D., & Christakis, N. A. (2012). Aspirin use and cardiovascular events in social networks. *Social Science & Medicine*, 74(7), 1125–1129. doi:10.1016/j.socscimed.2011.12.033 PubMed

TheBalanceCareers. (2018). Twitter User Statistics 2008 Through 2017. Retrieved 30 November 2018, from <https://www.thebalancecareers.com/twitter-statistics-2008-2009-2010-2011-3515899>

Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D. W., & Middleton, B. (2005). The Value of Health Care Information Exchange and Interoperability. *Health Affairs (Project Hope)*, 24(Suppl1), W5-10–W5-18. Advance online publication. doi:10.1377/hlthaff.W5.10 PubMed

Wennberg, J., & Gittelsohn, A. (1973). Small Area Variations in Health Care Delivery. *Science*, 182(4117), 1102–1108. doi:10.1126/science.182.4117.1102 PubMed

ADDITIONAL READING

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine*, 357(4), 370–379. doi:10.1056/NEJMs066082 PubMed

Christakis, N. A., & Fowler, J. H. (2008). The Collective Dynamics of Smoking in a Large Social Network. *The New England Journal of Medicine*, 358(21), 2249–2258. doi:10.1056/NEJMs0706154 PubMed

Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press.

Kim, D. A. (2015). *Social Networks and Health: From Epidemiology to Intervention*. Doctoral dissertation, Harvard University.

Morris, M. (1993). Epidemiology and Social Networks: Modeling Structured Diffusion. *Sociological Methods & Research*, 22(1), 99–126. doi:10.1177/0049124193022001005

Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2016). Official Journal of the European Union, L 119(4.5.2016).

Rumbold, J. M. M., & Pierscioneck, B. (2017). The Effect of the General Data Protection Regulation on Medical Research. *Journal of Medical Internet Research*, 19(2), e47. Advance online publication. doi:10.2196/jmir.7108 PubMed

Wilson, R. J. (2010). *Introduction to Graph Theory* (5th ed.). Pearson.

KEY TERMS AND DEFINITIONS

Data Protection: Process of protecting important information from corruption, compromise, or loss.

Information System: A software to collect, process, store, and disseminate information to support decision making, analysis, and visualization in an organization.

Medical Data: Health-related information that is associated with regular patient care.

Medical Network: A medical-based application of the principles of social networks.

Public Health: Health of the population as a whole.

Public Policy: Activities of governments acting directly or through delegation and influencing citizens' lives.

Social Network: A graph representing social interactions between individuals.

Ambiguities in the Privacy Policies of Common Health and Fitness Apps

15

Devjani Sen*Algonquin College, Canada***Rukhsana Ahmed** <https://orcid.org/0000-0003-0381-4491>*University at Albany, SUNY, USA*

INTRODUCTION

Mobile leisure, health, and wellness applications (apps) are ubiquitous. Research suggests that there are approximately 97,000 varieties of inexpensive and easy to use mobile health apps available in the market; at such a pace numbers are becoming outdated almost as soon as they are published (Privacy Clearinghouse, 2013). With approximately 320,000 of health and fitness apps in major app stores (Young, 2018), the question arises as to what happens to the sensitive data consumers enter into these apps, and what happens when these apps share data with advertisers and other third parties without the user's knowledge.

A growing topic of interest in both Canada and the U.S. concerns exactly what third parties can legally do with personal data. American law dictates that health insurance companies cannot discriminate based on a history of illness, specifically, severely restricting the dissemination and distribution of private health information without documented consent. However, while data held by a health plan, health care provider, or lab may be protected by the federal Health Insurance Portability and Accountability Act (HIPAA), legal scholars warn that if a patient is going to upload health or wellness data to a mobile application (app), it may not be covered by those laws (Rogers, 2014). Such legal ambiguities have implications for Canadian users of health and wellness apps, because many of these devices are based in the U.S., with the data being stored on U.S. servers and thus they may not conform to privacy requirements (Akkad, 2013). Clearly, such privacy concerns apply globally in any cases where personal data may be shared to third parties across two or more countries anywhere in the world.

There are some other important concerns with privacy and security issues related to mobile health and fitness applications (Huckvale et al. 2015; Rajindra et al. 2014). For example, personal apps collect all sorts of personal information like name, email address, age, height, weight, and in some cases detailed health information. When using such apps, many users may share a host of personal information and consequently make themselves targets to misuse of this information by unknown third parties. Moreover, according to Gralla et al. (2011), apps can gather the phone number and the unique ID number of each type of phone. In this way, personal information that apps gather about an end-user can be matched to these IDs, which means that ad networks can easily combine various pieces of information collected by multiple apps to build a sophisticated profile about a given end-user and thereby posing a major privacy risk to personal data. Therefore, un-informed decision by end-users raises important concerns regarding the ethics around sharing personal data gathered from health and fitness apps to third parties. To summarize, the issues raised above may be broken down to the following concerns:

DOI: 10.4018/978-1-7998-3479-3.ch127

- (1) ownership and veracity of sensitive data shared on personal apps
- (2) what end users really understand about the use of their data (what data are being collected and the specifics of how it may be used)
- (3) the ethics of sharing end-users' personal information and sharing it with third-parties

Despite the important role of informed consent in the creation of health and fitness mobile applications, the intersection of ethics and sharing of personal information is understudied and is an often-ignored topic during the creation of mobile apps. After reviewing the online privacy policies of a select set of mobile health and fitness apps, this chapter will conclude with a set of recommendations when designing privacy policies for the sharing of personal information collected from health and fitness apps.

BACKGROUND

Online privacy policies, which regulate the relationship between the user and the website with the purpose of limiting companies' legal liability during site use, are also employed by users to inform their understanding of the manners in which personal data are treated by companies. Despite their importance to users, however, studies suggest that these policies are often ignored (Angulo, Fischer-Hübner, Weastlund, & Pulls, 2012; Jensen & Potts, 2004; Kesan, Hayes, & Bashir, 2012; Tsai, Egelman, Cranor, & Acquisti, 2011). As pointed out by Steinfeld (2016), since agreeing to the terms of the policy is usually a prerequisite for subscribing to a website or a web service, users typically sign their consent almost automatically, so that these terms are rarely considered as reasons for joining or avoiding a given website.

Studies suggest that many apps do not have a privacy policy, or that apps do not grant users access to and control over personal information before users downloads and/or after using apps (Privacy Rights Clearinghouse, 2013). Nevertheless, having a privacy policy or providing a link to more information, is not enough to safeguard end-users' right to data privacy. Research suggests that innovative techniques must be developed to present information that was not possible with paper such that electronic formats do not simply mimic the metaphors of paper documents (Tansel, 2013). Research has shown that reading from electronic formats, such as the Web increases cognitive load and disorientation, and is consequently less efficient for learning (recognition as well as recall of information) than reading from paper documents (Eveland & Dunwoody, 2001). In their study of differences in cognitive processing when reading paper documents versus electronic documents, O'Hara and Sellen (1997) suggest a number of advantages to paper that need to be addressed in the design of digital systems. A chief advantage of paper, is the way it supports annotation while at the same time permitting quick and easy navigation, which in turn enables a reader to develop a sense of overall structure of the reading material. Furthermore, they learned that improvements in navigation and control of spatial layout of individual and multiple documents must also be supported in electronic documents. More specifically, studies have identified a number of cognitive factors which limit the comprehension of existing privacy policies, including complexity, legal language, and length (Angulo et al., 2012; Milne & Culnan, 2004; Nissenbaum, 2011; Tsai et al., 2011), use of vague terms and concepts (Anton et al., 2003), as well as design issues such as format and font size (Milne & Culnan, 2004).

IDENTIFYING PRIVACY RISKS IN HEALTH AND FITNESS APPLICATION PRIVACY POLICIES

15

Clearly, if privacy policies are not being read, there is a need for more readable policies that will be better understood and will more effectively inform users' decisions on whether it is prudent for them to join a given website service agreement. Indeed, given the varied educational levels and socio-economic backgrounds of online consumers that encompasses a wide range of the global population, it is critical that companies communicate effectively with customers through their online privacy policies (Wheatman & Ghiselli, 2014). With these considerations in mind, the purpose of the present chapter was to investigate the presence of important information and its presentation in plain language format in the privacy policies of health and fitness applications. Considering the limited scope of this chapter, the number of apps to examine was limited to four in order to properly execute a combined approach of using a checklist and discussing relevant excerpts from the selected privacy policies.

Health and fitness applications are software applications that offer health-related services on portable devices such as smartphones and tablet PCs. The apps examined in this chapter included: 1) Fitbit, which records data such as the number of steps walked, heart rate, quality of sleep, steps climbed, and other personal metrics; 2) JEFIT, a free app that records and logs fitness exercises, tracks workouts and progress, enables access to a large exercise database; 3) Nike Training Club, which gives the user a personal trainer, with more than 60 custom-built workouts from a Nike Professional Trainer as well as enables user to sync their results to social media sites such as Facebook to show friends how training is going; and 4) MyFitnessPal, a free calorie counter smartphone application and website, which helps to track calorie intake for exercise and diet and employs gamification elements to engage and motivate users to achieve their goals. In order to ensure that the health and fitness apps examined comprised popular apps with a wide user base, the selection criteria for each app was set at a minimum of one million downloads. Care was also taken to ensure that: a) each app had an androgynous user base, having an equal appeal for male and female users and b) two of the four apps in the sample included free apps considering such apps are expected to hold greater biases for selling material to third party sites. It is also important to note that the apps selected offer a diversity of core functions, including recording, logging and tracking personal metrics, diet, exercise and workout routines and progress, even working with certified personal trainers.

In order to assess the privacy policies of the selected four health and fitness apps, a checklist was used that comprised five privacy risk categories. These privacy risks were based on two questions: a) is important information missing to make informed decisions about the use of personal data? and b) is

Table 1. Checklist of Privacy Risks for Fitbit, JEFIT, Nike and MyFitnessPal Mobile Applications

Description of Privacy Risk	Fitbit	JEFIT	MyFitnessPal	Nike Training Club
Explicitly stated that under no circumstances will Personal Identifiers (PI) be shared	×	×	×	×
Use of cookies explicitly stated as well as what information is tracked and how it is used	✓	✓	✓	✓
Explicitly stated what sharing is done with third party sites and the security regarding links to other Sites	✓	✓	✓	✓
Location and use of personal data after account is deactivated	✓	×	×	×
Ownership of data is stated clearly when device is used by international users	✓	×	✓	✓

information being shared that might compromise the end-user's right to privacy of that information? The two authors of this chapter examined the online privacy policies of each selected app in search of important privacy risks. Based on their search, each author completed a separate checklist and compared their findings to reach an agreement of the presence or absence of each privacy risk category. Table 1 below outlines the results of the final checklist assessing the occurrence of each privacy risk examined.

The following section will discuss in further detail the subsections from the checklist from the assessment of the privacy policies of the four health and fitness applications.

Type of Data Collected Explicitly Explained and It Is Stated What Is Done with Personally Identified (PI) Data

Regarding personal identifiers, JEFIT states:

The information we collect in order to create your JEFIT.com account includes data such as your name and email address. The email address may be used to contact you with matters regarding JEFIT, such as a newsletter, notifications of new messages from other JEFIT members, or an announcement about a new product or service. We also collect some personal fitness information so that we can provide you with a personalized plan for successful weight loss. JEFIT also collects other profile data including but not limited to: personal interests, geographical location, gender, and age in order to assist our members find and communicate with each other. JEFIT collects personal information about you when you sign up for a JEFIT account. This information includes your age, metric preferences, and other personal fitness details. JEFIT also stores the data you enter about workouts, exercise, and anything you choose to track and keep records of.

When referring to non-personally-identifiable information, the policy states:

JEFIT also logs non-personally-identifiable information including IP address, profile information, aggregate user data, and browser type, from users and visitors to the site. This data is used to manage the website, track usage and improve the website services. This non-personally-identifiable information may be shared with third-parties to provide more relevant services and advertisements to members. User IP addresses are recorded for security and monitoring purposes.

From the above excerpts, it is not completely clear whether non-personally-identifiable information may be shared with third-parties. It is simply stated that personal information is stored and collected, without further explanation about the purpose of this storing and collecting. What does it really mean to use this data to provide the end-user with a personalized plan for successful weight loss? Does it mean sharing this information with third parties who can advise on how to best do this?

Further, in the section entitled Privacy Policy Change, JEFIT states:

When you use the JEFIT.com sites to provide personally identifiable information, you consent to the collection and use of that information in accordance with this privacy policy.

What does this statement convey? Again, the intended or planned use of the collected and stored personally identifiable information is not clearly communicated.

Finally, of the privacy policies of the four health and fitness applications explored, the most alarming ambiguity seems to be stated in the privacy policy of MyFitnessPal:

Except as otherwise provided in this Privacy Policy, and when You otherwise give permission or under the following circumstances, MyFitnessPal will not share Your Personal Information with third parties.

However, what is stated in subsection User Content, under Terms and Use, appears to conflict with (interestingly located in a poor visibility area of 19th bullet) MyFitnessPal's explicit statement regarding the PI data that the company can do whatever it chooses with the end-users personal data:

“User Content” is any content, materials or information (e.g., any text, information, photos, images, video, and other content and material, including nutritional information contributed to the Food Database) that You upload or post to, or transmit, display, perform or distribute by means of, the Services, whether in connection with Your use of Website, Application, or through the use of any Third Party Websites or Third Party Services or otherwise. You hereby grant MyFitnessPal and its officers, directors, employees, agents, affiliates, representatives, service providers, partners, sublicensees, successors, and assigns (collectively, the “MyFitnessPal Parties”) a perpetual, fully paid-up, worldwide, sublicensable, irrevocable, assignable license to copy, distribute, publish, transmit, publicly display or perform, edit, modify, translate, reformat and otherwise use User Content in connection with the operation of the Services or any other similar services or related business, in any medium now existing or later devised, including without limitation in advertising and publicity. You further agree that this license includes the right for the MyFitnessPal Parties to publish, display or otherwise use and make available your User Content and possibly your name and/or any user name of yours in connection with their exercise of the license granted under this section. You agree to waive, and hereby waive, any claims arising from or relating to the exercise by the MyFitnessPal Parties of the rights granted under this section, including without limitation any claims relating to your rights of personal privacy and publicity.

Regarding personal identifiers, Fitbit states the following:

We will never sell your data, and will only share personally identifiable data when you direct us to (or under the circumstances outlined in our Privacy Policy).

Although this information appears to be clear, it might not be understood the same way by different users. For example, many people might understand the above to mean that if they do not give explicit consent their data will not be shared. However, it could also be that directing data to be shared might occur in other more covert ways that most end-users are not aware of.

While NIKE does state in what ways PIs may be shared with other parties, they also explicitly state:

However, like other companies, NIKE cannot guarantee 100% the security or confidentiality of the information you provide to us.

Use of Cookies, What Information Is Tracked and How It Is Used

Most privacy policies of the health and fitness applications examined defined what is meant by cookies and what they do (how they track information) and give details about how to disable cookies. However,

many useful features of fitness apps warrant the use of entering in information that might be tracked by cookies. As Fitbit states:

When You Add Information to Your Account

You can customize your Fitbit experience by adding other types of information to your account, such as entering a food log or setting an alarm, personalizing your profile with photos, participating in discussion boards, or sending messages to your Fitbit friends. Whenever you add this type of data, we collect it and store it in your Fitbit account.

This statement raises the question of whether it is realistic for all cookies to be disabled. For example is it reasonable to expect a user to sign into a fitness app where they never enter information into food logs, set an alarm, participate in discussion boards, add photos to their profile, or interact with their friends who have this same app?

Sharing with Third Party Sites

On the whole, this section was well done by the four mobile application privacy policies examined. All apps stated that if end-users choose to share their information with Third Party Sites, they need to carefully read the privacy policies of these third party sites. This way, sharing of personal information with third party sites may likely happen with end-users' knowledge and based on their informed decision.

Location and Use of Personal Data After Account Is De-Activated

With the exception of Fitbit, it was not possible to clearly locate information on what happens to personal information once the user de-activates his or her account. Fitbit states:

You can deactivate your Fitbit account by contacting Customer Support. When you do, data that can identify you will be removed from the Fitbit Service, including but not limited to your email, name, photo(s), friends list and links to sites such as Facebook and Twitter. Backup copies of this data will be removed from our server based upon an automated schedule, which means it may persist in our archive for a short period. Fitbit may continue to use your de-identified data after you deactivate your account.

This statement raises two questions: 1) what happens to PIs during the short period of time referred to? and 2) what is the length of this period of time? Clearly more attention needs to be given to these important issues in terms of clarity and completeness of the information provided.

Ownership of Data Is Stated Clearly When Device Is Used by International Users

Only Fitbit clearly stated that the U.S. Privacy Policies do not apply to the sharing of personal data when using Fitbit devices outside the United States. Interestingly, although NIKE and MyFitnessPal also stated this information, it was missed entirely by one of the authors due to poor visibility. This may

be an important concern for international use of such devices by travelers from the U.S. in addition to end-users who reside in other countries.

RECOMMENDATIONS

Based on the review of the online privacy policies of four mobile health and fitness applications, and reflecting on the analysis of ambiguities and privacy risks in the privacy policies of those apps, the following recommendations are offered. Although these recommendations should be taken into consideration when designing privacy policies for the sharing of personal information collected from health and fitness apps, they can also be extended to the creation of privacy policies of mobile apps and web services in the areas of human resources, planning and project management, customer engagement, property management, education and training, among others.

1. It is suggested that personally identifiable information be stated clearly; PI and non-PI information should be clearly distinguished. It is also important to clearly state what will never be done with PI data. Moreover, this information should be made highly visible (within the first page) of any Privacy Policy/Terms and Conditions.
2. It was found that cookies were clearly defined in the privacy policies of the apps examined. However, given that information is tracked and used by cookies, the kinds of activities that enable activation of cookies might need to be stated more clearly.
3. Generally, when the end-user enters a Third Party Site, its privacy policies apply. While it is recommended that this guideline be stated clearly, the need to read and examine these third party site privacy policies should be highlighted as well. Furthermore, if Contests and Giveaways comprise Third Party sites, this information also needs to be highlighted; at some part of a Contest or Giveaway advertisement, it clearly needs to be stated that the privacy policy of this site is now in effect (vs. the privacy policy of the original site).
4. It is advised that information regarding the location and use of personal data after an account is disabled be stated clearly in the privacy policies. Not doing so may suggest that all data is automatically deleted upon deactivation of an account. Hence, it is recommended that this information be clearly presented under a subheading such as What Happens When Your Account is De-Activated or something similar to this.
5. It is proposed that the rightful owner of data be more clearly stated when devices are used by international users and this information be stated in a clearly visible manner.

FUTURE RESEARCH DIRECTIONS

This encyclopedia provides a comprehensive summary of research-based chapters on copious aspects confronting the field of information science and technology. Integrating ICTs into health care is complicated and it is a daunting process to create an ICT-driven health care system that is accessible, reliable, safe, and thus effective (Avtgis & Polack, 2013). Future research should proceed with caution when considering the design of health and fitness applications on mobile devices to help safeguard privacy when sharing personal information on health and fitness apps.

As explored in this chapter, mobile health and fitness applications raise many pertinent yet unanswered questions for health care ethics. Many end-users may automatically consider health and fitness apps to be beneficial; but, there are complications that may arise from neglecting privacy and security concerns when sharing personal information on these apps. Overall, ICTs help facilitate access to health information and thus can help to increase awareness which translates to patient participation in their own health care. There is an app for just about any health, fitness, and weight-loss resolution. Amidst this myriad health and fitness apps that are permeating lives all around the globe, an important question of data privacy looms large: what happens when advertisers and other third parties share this information without the user's knowledge? Questions like this and others that remain largely unexplored and that have only recently been asked by scholars should guide future investigations. Future studies should continue to investigate the complexities at the intersection of ethics and information sharing on health and fitness apps in order to tap into the potential of mobile apps to improve health behavior and health care delivery.

Health and fitness mobile application development should integrate the end-user experience perspective. Future research should privilege the end-user voice in determining ethics, privacy, and security of personal data shared on these apps, and in co-designing guidelines for data protection. Surveys, focus groups, and interviews can help gather important information regarding end-user awareness and comprehension of data privacy. Such data could yield valuable information to aid in designing simple and user-friendly privacy policies. Researchers across disciplines should continue to share good practices with regard to personal data use, sharing, and retention practices of health and fitness applications.

As the field of information science and technology continues to evolve, so do the available ICTs become integrated and may eventually be an integral part of the health care system. The end-users of mobile applications will have opportunities to influence developments in health information technologies that can lead to personalized tools to promote health and prevent disease. Yet, the intersections of ethics, privacy, security, and information sharing is understudied and is an often-ignored topic during the creation of mobile apps. In order to complement the research reported in this chapter, further exploration in the following areas is advocated for:

- A comprehensive review of privacy policies of the plethora of health and fitness applications that are available
- Full investigation, from the perspectives of both app developers and end-users, of ethical ambiguities in privacy policies across health and fitness applications
- An advancement of current understanding of how people with limited literacies – functional, health, digital, technology, and so on – use and apply health and fitness apps

CONCLUSION

Undoubtedly, the success of any emerging technology will depend to a large degree on the ability of people to use them effectively and efficiently. Today's health care system greatly depends on information and communication technologies. These advancements facilitate communication between patients, doctors, and other health care professionals. Through its review of topics such as the ethics of information sharing on apps and online privacy policies, this chapter helped to advance knowledge on the dangers of sharing personal information on mobile apps, which can jeopardize the rights to privacy of end-users of mobile apps. This chapter calls for multidisciplinary and multi-sectoral research partnerships to best meet the challenges of today's information technologies. Health and fitness mobile applications will not address

ethical and privacy concerns in and of itself. Researchers, practitioners, policy-makers, app developers, programmers, promoters, and users need to come together to debate and address these concerns.

ACKNOWLEDGMENT

An earlier version of this research was published in Khosrow-Pour, M. (2018). *Encyclopedia of Information Science and Technology* 4th edition.

REFERENCES

- Angulo, J., Fischer-Hübner, S., Weastlund, E., & Pulls, T. (2012). Towards usable privacy policy display and management. *Information Management & Computer Security*, 20(1), 4–17. doi:10.1108/09685221211219155
- Anton, A. I., Earp, J. B., & Carter, R. (2003). Precluding incongruous behavior by aligning software requirements with security and privacy policies. *Information and Software Technology*, 45(14), 967–977. doi:10.1016/S0950-5849(03)00095-8
- Avtgis, T. A., & Polack, E. P. (2013). Coming full circle in rural trauma: Chronicling the development and testing of communication systems in rural trauma networks. In R. Ahmed & B. R. Bates (Eds.), *Health communication and mass media: An integrated approach to policy and practice* (pp. 151–164). Gower Publishing Ltd.
- El Akkad, O. (2013, December 19). How insurers are turning to fitness apps to decide your health coverage. *The Globe and Mail*. Retrieved from <http://www.theglobeandmail.com/technology/how-insurers-are-turning-to-fitness-apps-to-decide-your-health-coverage/article16065068/>
- Eveland, W. P., & Dunwoody, S. (2001). User control and Structural Isomorphism or Disorientation and Cognitive Load? Learning from the Web versus Print. *Communication Research*, 28(1), 48–78.
- Fitbit. (2016). Fitbit privacy policy. Retrieved from <https://www.fitbit.com/ca/privacy>
- Gralla, P., Sacco, A., & Fass, R. (2011). Smartphone apps: Is your privacy protected? Retrieved from <http://www.computerworld.com/article/2509878/dataprivacy/smartphone-apps--is-your-privacy-protected-.html>
- Huckvale, K. T., Prieto, J., Tilney, M., Benghozi, P. J., & Car, J. (2015). Unaddressed privacy risks in accredited health and wellness apps: a Cross-sectional systematic assessment. Retrieved from: <https://bmcmmedicine.biomedcentral.com/articles/10.1186/s12916-015-0444-y>
- JEFIT. (2015). Privacy policy. Retrieved from <https://www.jefit.com/privacy-policy/>
- Jensen, C., & Potts, C. (2004). Privacy policies as decision-making tools: An Evaluation of online privacy notices. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, 471–478.
- Kalyani, M. (2013). Privacy Concerns with Mobile Health Apps. Retrieved from <https://spideroak.com/privacypost/cloud-security/privacy-concerns-with-mobile-health-apps-draft/>

Kesan, J., Hayes, C., & Bashir, M. (2012). Information privacy and data control in cloud computing: Consumer privacy choices, informed consent, and baseline protections to facilitate market transactions in the cloud. Illinois Program in Law, Behavior and Social Science, 11-20. Paper No. LBSS12-11. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2042638

Milne, G. R., & Culnan, M. J. (2004). Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing*, 18(3), 15–29. doi:10.1002/dir.20009

MyFitnessPal. (2013). Retrieved from http://www.myfitnesspal.com/account/terms_and_privacy?with_layout=true#privacy

Nike. (2016). Nike digital privacy policy. Retrieved from https://helpall.nike.com/app/answers/detail/a_id/52863/~nike-digital-privacy-policy

Nike. (2018). Nike digital privacy policy. Retrieved from: [https://agreementsservice.svs.nike.com/us/en_us/rest/agreement?agreementType=privacyPolicy&uxId=com.nike.nikegrind&country=US&language=en&requestType=redirect&rehttps %253A%252F%252F www.nikegrind.com%252F](https://agreementsservice.svs.nike.com/us/en_us/rest/agreement?agreementType=privacyPolicy&uxId=com.nike.nikegrind&country=US&language=en&requestType=redirect&rehttps%253A%252F%252Fwww.nikegrind.com%252F)

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, 140(4), 32–48. doi:10.1162/DAED_a_00113

O'Hara, K., & Sellen, A. (1997). A Comparison of reading paper and on-line documents. In *Proceedings of CHI '97 Conference on Computer Human Interaction*. ACM Press., doi:10.1145/258549.258787.

Privacy Rights Clearinghouse. (2013). Fact Sheet 39: Mobile health and fitness Apps: What are the privacy risks? Retrieved from <https://www.privacyrights.org/mobile-health-and-fitness-apps-what-are-privacy-risks>

Rajindra, A., Richards, D., & Scott, K. (2014). Security and Privacy Issues Related to the Use of Mobile Health Apps, *Proceedings of the 25th Australasian Conference on Information Systems*.

Rogers, J. (2014, September 16). Expert: Apple's health push poses big privacy challenges. Retrieved from <http://www.foxnews.com/tech/2014/09/16/expert-apples-health-push-poses-big-privacy-challenges/>

Steinfeld, N. (2016). "I agree to the terms and conditions": (How) do users read privacy policies online? An eye-tracking experiment. *Computers in Human Behavior*, 55(Part B), 992-1000.

Tsai, J. Y., Egelman, S., Cranor, L., & Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2), 254–268. doi:10.1287/isre.1090.0260

Wheatman, S., & Ghiselli, M. (2014). Privacy policies: How to communicate effectively with consumers. Retrieved from https://iapp.org/media/pdf/knowledge_center/IAPP_KMPrivacyPaper_FINAL.pdf

Young, L. (2018). There are thousands of health and fitness apps, and not much evidence they work: study Retrieved from <https://globalnews.ca/news/4215969/do-fitness-apps-work/>

ADDITIONAL READING

Gu, J., Xu, Y., Xu, H., Zhang, C., & Ling, H. (2016). Privacy concerns for mobile app download: An elaboration likelihood model perspective. *Decision Support Systems*, 95, 19–28. doi:10.1016/j.dss.2016.10.002

Huckvale, K., Prieto, J. T., Tilney, M., Benghozi, P.-J., & Car, J. (2015). Unaddressed privacy risks in accredited health and wellness apps: A cross-sectional systematic assessment. *BMC Medicine*, 13(1), 214. doi:10.1186/s12916-015-0444-y PubMed

O'Loughlin, K., Neary, M., Adkins, E. C., & Schueller, S. M. (2018). Reviewing the data security and privacy policies of mobile apps for depression. *Internet Interventions*, 15, 110–115. doi:10.1016/j.invent.2018.12.001 PubMed

Rowan, M., & Dehlinger, J. (2014). A privacy policy comparison of health and fitness related mobile applications. *Procedia Computer Science*, 37, 348–355. doi:10.1016/j.procs.2014.08.051

KEY TERMS AND DEFINITIONS

Ethics: The critical examination of the advantages and disadvantages when deciding upon the correct conduct involving a moral issue.

Information Security: The process of protecting the availability, integrity and privacy of information.

Mobile Applications: A term used to describe Internet applications that run on portable devices such as smartphones and other mobile devices to make it easier for users to access the Internet.

Mobile Health and Fitness Applications: Application programs that offer health-related services on portable devices such as smartphones and tablet computers.

Online Privacy Policy: A document, typically required by law, which regulates the relationship between the user and the website with the purpose of limiting companies' legal liability during site use.

Personal Identifiers Information (PIIs): Comprises information, that when considered alone, or in combination with data from other sources, may contribute to distinguish (identify) an individual.

Privacy: An individual's right to control how and to what extent information about him or her may be shared and acted upon by others.

Generosity in Healthcare Policy Under the Obama Administration: Reflecting Various Dimensions Focused on the ACA

Khadijeh Roya Rouzbehani

 <https://orcid.org/0000-0002-6045-048X>

University of Victoria, Canada

INTRODUCTION

The goal of the Patient Protection and Affordable Care Act (ACA) was to achieve nearly universal health insurance coverage in the United States through a combination of policies largely implemented in 2014 (Obama, 2016). Several recent studies, including Frean, Gruber, and Sommers (2017) and Courtemanche et al. (2017), have shown that the ACA led to gains in insurance coverage. In fact, Patient Protection and Affordable Care Act (ACA) of March 2010 was to achieve nearly universal health insurance coverage in the United States through a combination of insurance market reforms, mandates, subsidies, health insurance exchanges, and Medicaid expansions (Gruber, 2011). These major components of the ACA all took effect in 2014, with the Medicaid expansion being optional for states after a Supreme Court decision. Nowadays, the Affordable Care Act (ACA) constitutes an effective public policy alternative that remedies some of the previous problems that existed in America's healthcare system (Gholipour and Rouzbehani, 2016; Rouzbehani, 2017). The ACA seeks to accomplish several fundamental objectives namely reforming the private insurance market, expanding Medicaid to the working poor, and changing the way medical decisions are made in end-of-life situations. All these objectives are rooted in private choices and expectations of rational decision making and incentives (Silvers, 2013). The objectives of the ACA, when examined from the perspective laid out above, in terms of the maximization of net value achievement, could be perceived in the context of the rational resource allocation policy model (Dye, 1981; Birkland, 2005).

BACKGROUND

The Affordable Care Act (ACA) expanded Medicaid eligibility to persons earning up to 138% of the federal poverty level, as part of the largest expansion of coverage to nonelderly adults since the 1960s. Although the expansion was originally intended to be enacted nationally, a 2012 U.S. Supreme Court decision made it optional for states. A total of 24 states decided not to expand in 2014, which affected 6.7 million uninsured low-income adults who otherwise would have gained eligibility. Since 2014, an additional 5 states have implemented expansions, although 19 states still have not adopted the expansion as of January 2017.

DOI: 10.4018/978-1-7998-3479-3.ch128

FOCUS OF THE ARTICLE

In this article, the author will discuss reports reflecting empirical studies of various dimensions focused on the ACA. As readers will note, these studies ranged from investigations of the apparent gender differences in support for the policy (Lizottle, 2015), the impact of the policy on dependent coverage mandate on insurance premiums (Depen and Bailey, 2015), and the implications of the ACA expansion to young adults in terms of inpatient hospitalizations (Antwi, Moriya and Simon, 2015).

Some of the findings presented in these empirical studies confirmed the proposition that the ACA has led to improvement and thus exercised a positive impact on the healthcare system in the United States. However, other findings (see for example, Garson, 2000) also point to the fact that there are still greater tasks ahead in producing a healthcare system that is both affordable and that ensures quality of care for all population groups in the United States.

Research Hypothesis

The central hypothesis of this research is that the ACA led to major improvements on the previous health-care policy of the United States. This hypothesis is predicated on the fact that the previous system was inadequate. Thus, in the previous system there were many problems with the private insurance market. There was also a need to expand Medicaid to the working poor and change medical decisions during end-of-life situations. Further, the need existed in the previous system to close the insurance gaps for young adults. This need arose out of the fact that in the United States the transition from adolescence to young adulthood was associated with the collapse of one's insurance coverage (Antwan, Moriya and Simon, 2014). Overall, the previous healthcare policy did not meet the affordability test and could not ensure quality of care for all segments of the population regardless of socio-economic status, gender, race, or creed. To buttress the hypothesis that the previous system did not meet the affordability test, Herzlinger (2010) noted that public healthcare spending in the United States threatened the economic welfare of the country because it was a poor value proposition in relation to its cost. Garson (2000) for his part, acknowledged the fact that to make the system more viable there was a need for more savings. The author proposed a variety of tools that would help accomplish this task (pp. 1048-52).

These are just a few of the many challenges that the previous healthcare policy could not meet. On the heels of these problems, hence, a new policy framework was set in motion with the passage of the ACA. In this paper, the efficacy of the ACA using the rational resource allocation model as a conceptual framework is assessed. Other theoretical approaches are also used to provide clarity on how political and institutional variables impinge upon policy development in the healthcare field (Gholipour and Rouzbehani, 2016).

Methodological Framework

Alternative approaches to policy formation are often governed by various policy models (Dye, 1981; Arnold, 1990; Mann, 1991; Birkland, 2005; Anderson, 2003; Kingdom, 2003). This researcher argues that sometimes even if government decides not to do anything in relation to a perceived policy problem, such decision may also be governed by a policy model (Birkland, 2005). Thus, policy models enable us to clarify our thinking about government, politics, and the policy process (Dye, 1981; Kingdom, 2003). They also allow us to make sense of the motivations of policy actors and consequences of policy alternatives (Dye, 1981, p. 17). For purposes of this discussion, I adopt the rational resource allocation policy

model. This model is being used as a theoretical and conceptual framework to justify the ACA. And by extension, to support the above stipulated hypothesis. Dye (1981) states that a rational policy approach is one that is designed to maximize what he refers to as “net value achievement.” Net value achievement is defined as a situation in which any sacrifice in one or more values required by a policy is “more than compensated for by the attainment of other values.” (p.27).

To arrive at this policy equation, the author also stipulates as a precondition that all relevant values to a society are known. As a matter of interpretation, I would argue that this definition of rationality is akin to efficiency—where efficiency is a ratio between valued inputs and valued outputs (Dye, 1981, p.27). The efficiency standard suggests that a policy is rational when it is most efficient meaning the ratio of the values it achieves and the values it sacrifices is positive and thus higher than any other positive alternatives (p.27). Dye’s (1981, p.27) idea of efficiency as he states involves “the calculation of all social, political, and economic values sacrificed or achieved by a public policy, not just those that can be measured by quantitative symbols.”

I also adopted a plethora of other policy models to complement the central conceptual and theoretical framework (Dye, 1981). These models include group theory, incrementalism and institutionalism. Group theory was used for example, to help account for the interaction and banding among groups of individuals with common interests. The model of incrementalism was also used to explain why the ACA was preferred over other radical alternatives. Here is how Kingdom (2003) described the model of incrementalism:

“Partly in response to writings which imply that a rational-comprehensive model either is or should be used in governmental policy making, Charles Lindblom and others developed their description and defense of an incremental approach. Instead of beginning consideration of each program or issue afresh, decision makers take what they are currently doing as given, and make small, incremental, marginal adjustments in that current behavior...The result is that policy changes very gradually, in small steps (p.79).”

The above definition aptly defines incrementalism as operationalized in this study. Similarly, the model of institutionalism was also used to shed light on the role played by governmental institutions in both promoting and resisting the ACA.

The Rationale of the Affordable Care Act

Series of empirical researches on the ACA have been conducted. What forms the core of this section are presentations of their findings on the ACA. The section also relies on findings found in reports and evaluative studies of the ACA. The analytical objective here is to show why the ACA became necessary to replace an existing policy alternative that was problematic in many ways. The findings presented here covered various grounds in terms of positing a rationale for the new policy—economic, political, efficiency, humanitarian and institutional.

Thus, Lizette (2015) in a study on the gender gap in healthcare policy preferences posits that there is a constant gender gap in healthcare policy preferences as such that women are more likely to support the Affordable Healthcare ACA than men. The study investigates two explanations for this outcome. The results show that the gap is connected to a larger debate in political behavior in terms of whether self-interest versus symbolic reasons drives public opinion. The author puts forward two central hypotheses: a) a humanitarian hypothesis and, b) an economic hypothesis. The study concludes that there is support for both hypotheses, and each partially mediates the gender gap in support for the ACA.

Another study conducted on the ACA found that the employers' mandate has led to a 2.5-2.8 percent increase in premiums for health insurance plans that cover children, relative to single-coverage plans. The study concludes that employers did not pass on the entire increase in premiums to employees through higher required contributions. This study is not unique as its findings are confirmed in other studies (see also Herman, 2015; Antwi, Moriya and Simon, 2015).

Moriya and Simon (2015) investigated the implications of the ACA expansion to young adults in terms of inpatient hospitalizations. The study focused on mental health related inpatient care and concluded that compared to those aged 27-29 years, treated young adults aged 19-25 years increased their inpatient visits by 3.5 percent, while mental illness visits increased by 9.0 percent. The study also found that the prevalence of insurance among hospitalized young adults decreased by 12.5 percent.

The broader implication of this study is that allowing young adults to remain on their parents' health insurance until they turned 26 years old was a rational approach in the context of the chosen conceptual framework. For example, the issue of mental health among young adults is a vexing one in an era of mass shootings, especially in the context of alternative approaches to controlling gun related mass shootings. Hence, any healthcare policy impact that increases early diagnoses and screening of mental health patients that may be prone to violence increases net value achievement from a societal and public health perspective.

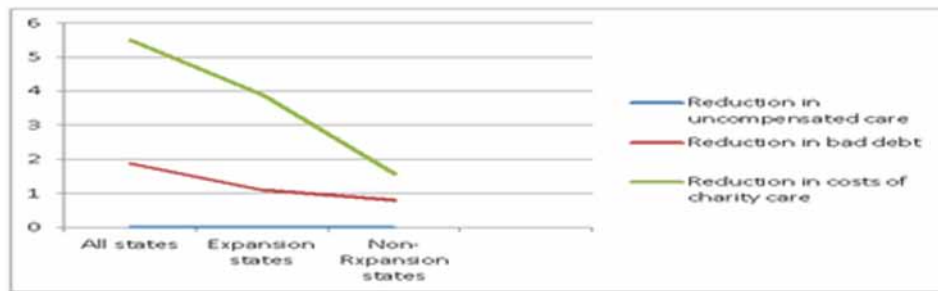
The ACA was based on Massachusetts' healthcare reform, which preceded it. In effect, the Massachusetts' healthcare reform was a model for the ACA. One thing that the Massachusetts' healthcare reform shows is its ability to dissolve some of the ethnic and racial gaps that exist in access to healthcare. Maxwell et al (2011) conducted a mixed methods research to assess the extent to which it decreased disparities in insurance coverage and access to healthcare. The authors concluded that although disparities remain, there was an increase in insurance coverage among Hispanics. The research findings showed that increase in health insurance for Hispanics more than double that experienced by non-Hispanic whites (p. 1415). The paper also pointed out that language and other cultural factors were significant obstacles for Hispanics. It was recorded that only 66 percent of Hispanics with limited proficiency in English were insured (p.1415). The authors further reported that by 2009, one-third of Spanish-speaking Hispanics did not have a personal provider and 26 percent reported seeing a doctor because of cost. To sum up, these results were revealing as they confirmed a secondary goal of health care reform in Massachusetts, which was to reduce persisting ethnic and racial disparities in coverage (p.1415).

On the question of ending segregated behavioral healthcare, which was a feature of the previous system, McLellan and Woodworth (2014) argued that this practice was scheduled to change. Under the ACA, healthcare providers were mandated to provide prevention, screening, and interventions for the full spectrum of substance use disorders, not just addiction (p.542). Further, the ACA was also meant to make it possible for healthcare plans to carve in behavioral healthcare under their major medical benefit with the aim of improving total health outcomes (p.542).

In a study conducted by Huntoon, McClunney and Scannell (2011) it was found that 94.8 percent of medical students surveyed agreed that the pervious healthcare system needed to be changed. Also, 80 percent of those surveyed indicated that they supported the ACA and 78.3 percent indicated that they did not think that the reform efforts under the ACA had gone far enough. Meanwhile, a majority of those surveyed (58.8 percent) also said that they opposed repeal of the new healthcare law, while only 15.0 percent opposed repeal of the ACA. This finding was almost counterintuitive at the time given the fact that the ACA itself was very unpopular. The table below is just one among several pieces of empirical evidence that proves that Medicaid expansion under the rubric of the Affordable Care Act has yielded positive results.

Figure 1. Estimated reduction in uncompensated care costs in 2014 as a result of marketplace coverage and Medicaid expansion (in billions of dollars).

Source: *U.S. News and World Report*, 2015, p.1.



Source: *U.S. News and World Report*, 2015, p.1.

The figure(1) shows some differences in outcomes among states that accepted the implementation of Medicaid expansion (expansion states) and those who refused (non-expansion states) the implementation of this crucial aspect of the policy. Thus, while expansion states recorded reductions in uncompensated care cost amounting to 26 percent or \$5.0 billion dollars, those who refused to implement the provision recorded only 16 percent reduction in uncompensated care cost, which amounted to \$2.4 billion dollars. The next section examines the political and institutional forces that impinged upon the ACA's conceptualization, justification, and enactment into law and public policy.

Institutional, Ideological, and Political Forces

Like all policy alternatives, there were specific political and institutional imperatives that led to the enactment of the ACA (Hayes, 2007). Invariably, these processes have also influenced the pattern of implementation of the policy. Thus, to fully appreciate these variables in a nuanced fashion, and as mentioned earlier, this researcher adopted a plethora of other policy models to reinforce the central conceptual and theoretical framework (Dye, 1981). These models include group theory, incrementalism and institutionalism. Group theory is applied for example, to help account for the interaction and banding among groups of individuals with common interests (pp.21-4). This researcher also used incrementalism to explain why the ACA was preferred over other radical alternatives (pp.31-3). Similarly, the model of institutionalism was operationalized to shed light on the role played by governmental institutions and partisan interests in both promoting and resisting the ACA (pp.18-21). In applying the model of institutionalism, this researcher draws on the wisdom of Jacob Hacker and Paul Pierson's (cited in Riby, Clark and Pelika, 2014) when they intimated that:

“At the heart of politics is the exercise of authority. Winners of political contests are positioned to use the control of the coercive power of the state to impose their preferences on losers through public policies. The stakes are often vast, involving hundreds of billions of dollars and, literally, matters of life and death. Policies are not a side-show. For major political actors, especially the most well-resourced and enduring, they are the central object of contestation. Elections are just one means to the end of exercising authority through policy. Because policy is the central prize, it deserves a central position in the analysis of political systems—a position it has lost in modern political science, at a considerable cost to our understanding of politics (p.59).”

During President Barrack Obama's historic campaign to become president, he put forward a series of policy proposals to tackle most of the challenging issues that faced the nation. One of those proposals was his healthcare plan that he promised would revolutionize the public healthcare system in the United States. The President was committed to seeing through the implementation of his healthcare plan despite the exigencies of restoring the health of the United States economy. Before his ascension, the economy had experienced severe shocks due to the collapse of the housing market. In fact, one would argue that at this time the country was at a precipice---the verge of a second great depression.

Pressures also mounted from many immigration reform support groups that the president did something about reforming the immigration system. However, in the midst of all these pressures from many of his core groups of supporters, President Obama decided to fulfill his promise to reform a healthcare system that many regarded as underperforming, broken and costly. The main opposing party, the Republicans, was in combat mode, ready and prepared to resist changes to the existing healthcare system that did not confirm to their ideas and philosophies that they thought should govern healthcare reform.

The Republicans fiercely argued that implementing the ACA would lead to a federal government takeover of the healthcare system. The irony in the situation is the fact that the main provisions of the ACA were modeled after Republican initiatives that initially had broad bipartisan support (Lanford and Quandagno, 2015, p. 17). After President Obama signed the ACA into law, 26 states immediately challenged key provisions of the law by filing lawsuits (Lanford and Quandagno, 2015). In 2012, the Supreme Court ruled in support of most provisions. However, the Court also ruled that the mandatory Medicaid expansion was unconstitutional, thereby making the implementation of this provision optional for the states.

President Obama and the Democrats in Congress expanded great political capital to achieve what could only be described as a compromise. Indeed, the ACA was merely a compromise in the face of fierce opposition to address the failures of the healthcare system to provide affordable healthcare to millions of Americans. As a result the single payer-approach was taken off the table (The Nation, 2013). The fact that the much popular public option that would have established the foundation for a single-payer system was never proposed signaled an incrementalist turn in the market-based approach of President Obama and the Democrats (Dye, 1981; Anderson, 2003). In effect, the ACA would end up being just an attempt at incremental improvement to the existing healthcare system. But as mentioned above, the Republican Party remained hostile to the policy despite this compromised policy outcome (see Rigby, Clark & Pelika, 2014).

SOLUTIONS AND RECOMMENDATIONS

Lanford and Quandagno (2015) examined factors explaining Medicaid expansion and concluded that there were several predictors, including states' Medicaid policy legacy, providers' influence, racial resentment, and conservative ideology. Based on these findings, the authors concluded that enactment of legislation is only a part of the process of policy formation. The authors suggested that implementation and state dynamics are equally salient variables.

Lanford and Quandagno (2015, p. 8) also found that each of their measures of partisanship was highly positively correlated with movement toward Medicaid expansion statistically ($p < .001$). Thus, what the authors referred to as Democratic power was said to have a strong positive influence on movement toward Medicaid expansion, and Republican power had recorded a strong negative influence (p.8). The study also found that in states where people identified more as liberals, there was movement closer to

Medicaid expansion than states where people identified more as conservatives ($r = -.58$; $p < .001$) (p.12). Racial resentment in the study was found to also be more closely linked to Medicaid expansion decisions (p.12). Overall, the pattern of political behavior demonstrated by both parties with regards to Medicaid expansion and other provisions of the ACA is broadly consistent with David Truman's (see Dye, 1981, p. 21) conceptions of the group theoretic approach to public policy.

The issue of Medicaid expansion is at the core of the various factors that would make the ACA a viable policy alternative. Indeed, Medicaid expansion is as important as other provisions of the ACA, such as the abolition of preexisting conditions, the health insurance mandate, states' establishment of health insurance exchanges etc. Thus, contrary to initial expectations, over half of the states have refused to create their own health insurance exchanges (Hall, 2014).

More than half the states have refused to expand Medicaid, decisions that are usually based on party affiliation. Thus, all the states, as observed in studies mentioned above, that have refused to expand Medicaid are usually led by Republican governors or legislative majorities (p. 1459). The problem here is that states that run their own insurance exchanges have local control over the very kind of important policy and regulatory matters that conservatives complained the federal government usurps (p. 1459).

From a policy implementation standpoint, these are challenges that have confronted the full realization of the policy goals of the ACA. In addition, one would argue that the Supreme Court ruling in *National Federation of Independent Business v. Sibelius*, that held that states are allowed to opt out of Medicaid expansion, without sacrificing their current Medicaid funding, has undermined a major plank of the ACA. Critics on the left of the Democratic Party have also argued that the fact that the public option was not pursued means the ACA did not go far enough in instituting public health policy reform. These issues are currently playing in the Democratic Party with the rise of the Bennie Sanders' campaign for President of the United States.

Other critics have pointed at the absence of relevant provisions that would address the lack of price and quality transparency in the system (Herzlinger, 2010). Yet other politicians in the Democratic Party have called for the American healthcare system to assume features of the Canadian, British, and other healthcare systems in developed capitalist countries. In fact, it should be pointed out that the United States and Canada had similar healthcare systems in the early 1960s. It was not until in 1966, with the passage of the Medical Care Act (MCA) that Canada started on the path of what would become a truly universal system of healthcare (Roberts and Liu, 2014, p. 764).

FUTURE RESEARCH DIRECTIONS

A promising research can focus on a systematic effort to collect and "map" information on existing task-sharing approaches in rural and otherwise underserved settings in the United States and abroad using a mixed methods approach.

CONCLUSION

The ACA put in place profound changes in the way health insurance markets operate in the US and made coverage available for millions of uninsured or underinsured Americans. This paper posited that the Affordable Care Act (ACA) constitutes an effective public policy alternative that remedies some of the previous problems that existed in America's healthcare system. The central hypothesis put forward

in this paper is that the ACA is an improved alternative to the previous policy regime on healthcare. This hypothesis is supported by the rational resource allocation model. Empirical evidence reflected in prior studies was also adduced to support the hypothesis. The research concludes that further academic studies are required to go beyond current heuristic frameworks to deepen our understanding of how the ACA could be improved. It is suggested that understanding this unique challenge in the field of American public policy could only be achieved by effectively gauging some of the dilemmas and intricacies associated with evolving institutional and interest group politics.

REFERENCES

- Anderson, J. E. (2003). *Public policymaking*. Houghton Mifflin Company.
- Antwi, Y. A., Moriya, A. S., & Simon, K. L. (2015). Access to health insurance and the use of inpatient medical care: Evidence from the Affordable Care Act young adult mandate. *Journal of Health Economics*, 39, 171–187. doi:10.1016/j.jhealeco.2014.11.007 PMID:25544401
- Birkland, T. A. (2005). *An introduction to the policy process: Theories, concepts, and models of public policy making*. M.E. Sharpe.
- Courtemanche, C., Marton, J., Ukert, B., Yelowitz, A., & Zapata, D. (2017). Early impacts of the Affordable Care Act on health insurance coverage in Medicaid expansion and non-expansion states. *Journal of Policy Analysis and Management*, 36(1), 178–210. doi:10.1002/pam.21961 PMID:27992151
- Depen, B., & Bailey, J. (2015). Did the Affordable Care Act dependent coverage mandate increase premiums? *Journal of Health Economics*, 41, 1–14. doi:10.1016/j.jhealeco.2015.01.004 PMID:25645977
- Dye, T. R. (1981). *Understanding public policy*. Prentice Hall, Inc.
- Frean, M., Gruber, J., & Sommers, B. D. (2017). Premium subsidies, the mandate, and Medicaid expansion: Coverage effects of the Affordable Care Act. *Journal of Health Economics*, 53, 72–86. doi:10.1016/j.jhealeco.2017.02.004 PMID:28319791
- Gholipour, R., & Rouzbehani, Kh. (2016). *Social, Economic and Political Perspectives on Public Health Policy Making*. IGI Publishing House. doi:10.4018/978-1-4666-9944-1
- Gruber, J. (2011). *The impacts of the Affordable Care Act: How reasonable are the projections?* National Bureau of Economic Research, Working paper no. 17168.
- Hall, M. A. (2014). States' decisions not to expand Medicaid. *North Carolina Law Review*, 92(5), 1459–1473.
- Hayes, M. T. (2007). Policy characteristics, patterns of politics, and the minimum wage: Toward a typology of redistributive policies. *Policy Studies Journal: the Journal of the Policy Studies Organization*, 35(3), 465–480. doi:10.1111/j.1541-0072.2007.00233.x
- Herman, B. (2015). Employers mull impact of new reforms law coverage mandate. *Journal of Modern Healthcare*, 45(1), 4.
- Herzlinger, R. E. (2010). Healthcare reform and its implications for the U.S. economy. *Business Horizons*, 53(2), 105–117. doi:10.1016/j.bushor.2009.10.002

Kingdom, J. W. (2003). *Agendas, alternatives, and public policies*. Longman.

Lanford, D., & Quandagno, J. (2015). Implementing Obamacare: The politics of Medicaid expansion under the Affordable Care Act of 2010. *Sociological Perspectives*, 1–21.

Lizotte, M. (2015). Investigating women's greater support of the Affordable Care ACA. *The Social Science Journal*, 52(2), 102–111.

Obama, B. (2016). United States health reform: Progress to date and next steps. *Journal of the American Medical Association*, 316(5), 525–532. doi:10.1001/jama.2016.9797 PMID:27400401

Rigby, E., Clark, J. H., & Pelika, S. (2014). Party politics and enactment of “Obamacare”: A policy-centered analysis of minority party involvement. *Journal of Health Politics, Policy and Law*, 39(1), 57–95. doi:10.1215/03616878-2395181 PMID:24193613

Roberts, C. W., & Liu, H. (2014). On the cultural foundations for universal healthcare: Implications from late 20th century U.S. and Canadian health-related discourse. *Journal of Communication*, 64(4), 764–784. doi:10.1111/jcom.12079

Rouzbehani, Kh. (2017). Health policy implementation: Moving beyond its barriers in US. In *Handbook of Research on Healthcare Administration and Management*. IGI Global.

Silvers, J. B. (2013). The Affordable Healthcare Act: Objectives and likely results in an imperfect world. *Annals of Family Medicine*, 11(5), 402–405. doi:10.1370/afm.1567 PMID:24019270

The Nation. (2013). *Why the left should defend Obamacare*. <http://www.thenation.com/article/why-the-left-should-defend-Obamacare/>

ADDITIONAL READING

Claxton, G., Rae, M., Panchal, N., Damico, A., Whitmore, H., Bostick, N., & Kenward, K. (2013). Health Benefits in 2013: Moderate Premium Increases in Employer-Sponsored Plans. *Health Affairs (Project Hope)*, 32(9), 1667–1676. doi:10.1377/hlthaff.2013.0644 PMID:23962411

Ku, L. (2009). Health Insurance Coverage and Medical Expenditures of Immigrants and Native Born Citizens in the United States. *American Journal of Public Health*, 99(7), 1322–1328. doi:10.2105/AJPH.2008.144733 PMID:19443829

Manning, W. G., Basu, A., & Mullahy, J. (2005). Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data. *Journal of Health Economics*, 24(3), 465–488. doi:10.1016/j.jhealeco.2004.09.011 PMID:15811539

Merluzzi, T. V., & Nairn, R. C. (1999). Adulthood and Aging: Transitions in Health and Health Cognition. In T. L. Whitman, T. V. Merluzzi, & R. D. White (Eds.), *Life-Span Perspective on Health and Illness* (pp. 189–206). Erlbaum.

Mulcahy, A., Harris, K., Finegold, K., Kellermann, A., Edelman, L., & Sommers, B. D. (2013). Insurance Coverage of Emergency Care for Young Adults under Health Reform. *The New England Journal of Medicine*, 368(22), 2105–2112. doi:10.1056/NEJMsa1212779 PMID:23718165

KEY TERMS AND DEFINITIONS

ACA(Affordable Care Act): Contains comprehensive health insurance reforms and includes tax provisions that affect individuals, families, businesses, insurers, tax-exempt organizations and government entities.

Healthcare: The maintenance and improvement of physical and mental health, especially through the provision of medical services.

Insurance Market: The insurance market is simply the buying and selling of insurance.

MCA (Medical Care Act): The federal government passed the Medical Care Act in 1966, which offered to reimburse, or cost share, one-half of provincial and territorial costs for medical services provided by a doctor outside hospitals.

Medicaid: It is a health care program that assists low-income families or individuals in paying for doctor visits, hospital stays, long-term medical, custodial care costs and more. Medicaid is a joint program, funded primarily by the federal government and run at the state level, where coverage may vary.

Obamacare: An informal term for a federal law intended to improve access to health insurance for US citizens.

Policy Model: The rational model of decision-making is a process for making sound decisions in policy-making in the public sector.

Islamic Medical Ethics in Brunei Darussalam Hospitals

Syazana Fauzi

Universiti Brunei Darussalam, Brunei

INTRODUCTION

Today, one of the most pressing controversial debates pertains to the ethics surrounding the medical field. Due to the effervescent technological advancement, the lines between what is ethical and unethical are obscured. Thus, there is a necessity to draw a fine line between what is deemed ethical and unethical. In Muslim societies, any issues or concerns, including those relating to medical ethics, are addressed, and often resolved by issuing *fatāwā* (religious rulings), derived from the learned Muslim scholars' *ijtihad* (reasonings). These *ijtihad* stemmed from two principal knowledge sources, the Qur'ān and the Sunnah of Prophet Muhammad ﷺ. To clarify, medical ethical issues are essentially concerns faced in the medical field in general; however, they are discussed from an Islamic perspective in this chapter. The question that poses itself now is: what exactly is Islamic medical ethics? More specifically, how does it differ from the Western opinion of medical ethics? To grasp the difference between Western and Islamic ethics, particularly in a medical environment, it is important to understand that the connotation of ethics and law in the West is distinctive than that of ethics and law in Islam. What is generally meant by ethics here, according to the *Oxford English Dictionary*, is a set of rules that governs a person's behaviour, or conduct, activity, or action, based on his moral values. From a Western perspective, medical ethics are moral principles that apply values and judgements to the field of medicine, including clinical practice and scientific research (Beauchamp & James, 2013).

Medical ethics in the West functions as a guideline for medical professionals and scientists towards protecting and defending human dignity and rights of the patients (World Medical Association Declaration on the Right of the Patient, 1981), whereas for Muslims, both practitioners and patients are obligated to adhere to medical ethics as defined by Islam. Simply put, there are specific guidelines on how a Muslim should behave and what he should undertake for certain duties in a personal or medical situation. For example, it is imperative for a Muslim patient in a critical condition to practise patience, or *ṣabr*, and contentment, or *riḍā*, (two ethical manners encouraged by Islam), and to pronounce the *shahādah* prior to death. Thus, the moral principles of medical ethics in Islam are based on the *Sharī'ah*. Furthermore, what may seem like a moral issue in the West may actually be ethical and legal issues in Islam. For example, in cases of induced abortion. In some Western societies, a non-Muslim woman may have a right to abort her foetus without facing legal consequences. However, if a Muslim woman aborted her foetus, then such an act is illegal under the *Sharī'ah*, as agreed upon by the International Islamic *Fiqh* Academy. Therefore, the *Sharī'ah* sets ethical principles by which Muslim medical practitioners, researchers, as well as, patients, are guided by. In essence, what makes medical ethics "Islamic" is that these ethics are based on Islamic principles. These medical ethics and principles are extracted from the Qur'ān and the Sunnah, explained as well by the contemporary *fatāwā*, particularly for the purposes of this research, those issued by Brunei Darussalam's State Mufti.

DOI: 10.4018/978-1-7998-3479-3.ch129

The *fatāwā* of Brunei Darussalam's State Mufti are selected as this research exclusively looks at the context of Brunei Darussalam. Brunei Darussalam is a small Islamic monarchy, located on the north coast of the Borneo island in South-east Asia. On the first of January 1984, when Brunei Darussalam gained independence, Sultan Haji Hassanal Bolkiah officially proclaimed the concept of 'Melayu Islam Beraja' ('Malay Islamic Monarch' in English) as the country's national philosophy. Although Islam has been Brunei Darussalam's official religion since the declaration of the 1959 Constitution, and Islam has been long practised in the country since the first Sultan converted to Islam in the early 1360s (Elisseeff, 2000), most, if not all, of the medical practices and ethics implemented in Brunei Darussalam were not influenced by Islamic principles (*Brunei Annual Report 1983–1984*, p. 242). Even though Brunei Darussalam appoints an official State Mufti, who provides religious verdicts on various issues, including several ethical issues relating to healthcare, Brunei Darussalam's Ministry of Health implements the Good Medical and Dental Practice guideline, as part of their medical ethics. All practising doctors and dentists in Brunei Darussalam must adhere to this set of ethical guidelines, formulated by the Brunei Medical Board and adapted from the Good Medical Practice (2006), with the permission of the General Medical Council, United Kingdom. The guideline not only addresses doctors and dentists, but it also informs the public on what they can expect from the medical professionals.

Because of this separation of healthcare and religion, one may assume that the way medical practitioners handle medical cases might not coincide with the State Mufti's *fatāwā* regarding certain ethical issues. Thus, this research aims to analyse whether Brunei Darussalam's health professionals handle specific medical ethical cases in a manner that is congruent or incongruent to the State Mufti's *fatāwā*. In order to achieve this aim, this research conducts semi-structured e-mail interviews that were sent to a non-random sample of three hospital nurses working under relevant departments, such as the Operation Theatre Department, as well as, the Obstetrics and Gynaecology Department. All interviewees received the same set of questions, regarding how they typically handle cases of abortion, euthanasia, and organ transplantation (see Appendix 1). As some of the interviewees' responses overlap or are similar to each other, this research only presents their main ideas or arguments. As for Brunei Darussalam's State Mufti's religious opinions, relevant books discussing his *fatāwā* concerning certain medical ethical questions will also be utilised in this chapter. Analysing and comparing these qualitative data reveals the level of congruency between the State Mufti's *fatāwā* and the real-life medical practices in Brunei Darussalam hospitals. The findings for this research will contribute to filling some organisational knowledge gap, pertaining to how health institutions address some current controversial medical ethical issues from an Islamic perspective, particularly in Brunei Darussalam, a nation that is undeniably understudied as a case study, and one that claims to be an Islamic system of governance, as Islam is one of the principal components of its socio-political schema. Having considered the significance of this research, the next section reviews the various literatures on Islamic medical ethics.

BACKGROUND

Before delving into the literature review, the term Islamic medical ethics must be clearly defined first. For the purposes of this research, Islamic medical ethics refer to the application of Islamic principles to guide ethical or moral behaviour related to the medical and scientific fields, which specifically deal with human life (Shomali, 2008). There are many issues pertaining to ethics in medical healthcare. However, only three contentious ones will be discussed here: (a) abortion – on the right to abort a viable foetus, (b) euthanasia – on the right to live or die, and (c) organ transplantation – on the right to donate or ac-

cept an organ. Although medical ethics and principles in Islam are well-established as they are based on the Qur'ān and the Sunnah, research on their relevant application to the contemporary world has only emerged recently. For instance, Gatrad and Sheikh (2001) co-wrote a paper called 'Medical Ethics and Islam: Principles and Practice', in which they approached the subject matter by first defining medical ethics and providing a typology of medical ethics. Several ethical issues from an Islamic perspective are also covered in the paper. Examples of ethical issues include, but are not limited to, genetic manipulation, assisted conception, adoption, autonomy, organ transplantation, end of life, and brain death. The authors highlighted three sources of the *Sharī'ah*: the first two are the Qur'ān and the Sunnah, as the primary sources. And the third, but secondary, source is *ijtihād*, that is, the deductive reasoning, or interpretation, based on one's understanding of the *Sharī'ah*. Gatrad and Sheikh (2001) mentioned that these three sources contribute to the notion of Islamic medical ethics.

Significantly, they also included the *maqāṣid* as the guiding principles of Islamic Law. It is important to appreciate the *maqāṣid*, or objectives of the *Sharī'ah*, as they have helped resolve some of today's ethical and moral dilemmas by establishing the parameters of Islamic medical ethics. In other words, for a medical decision, or action, to be considered as ethical, such action must be considered in light of at least one *maqṣad*. For example, the procedure of in vitro fertilisation (IVF), for legally married couples recognised by the *Sharī'ah* who have difficulty in conceiving a child, would align with the fourth *maqṣad* on the preservation of progeny (Al-Bar & Chamsi-Pasha, 2015, pp. 173–186). Essentially, the three sources highlighted by Gatrad and Sheikh (2001), namely the Qur'ān, the Sunnah, and *ijtihād*, act as the guiding principles to formulate appropriate ethical rulings, according to time, place, and circumstances. Although their paper provides a good outline and serves as a theoretical framework for addressing contemporary ethical issues in scientific research, as well as, in the practice of clinical medicine, it did not discuss the practical application of Islamic medical ethics in the current healthcare field, especially in hospitals.

Another paper pertinent to this discussion is Padela (2007)'s 'Islamic Medical Ethics: A Primer', which deliberates upon the essence of ethics within the Islamic intellectual and juridical discourses. Within the Islamic intellectual dialogue, the relation between *adab* (appropriate manner or conduct) and *akhlaq* (character) is discussed, in which it is believed that *adab* literature intends to refine *akhlaq*. Alternatively, the paper explains that the ethical codes within the Islamic juridical discussion are often derived from the *Sharī'ah* and Islamic *fiqh* (jurisprudential understanding). Similar to Gatrad and Sheikh (2001)'s paper, Padela (2007) also explains the three different sources of the *Sharī'ah*, which are: the Qur'ān, the Sunnah, and, *ijtihād*. The Islamic codes on medical ethics are derived from these three sources. Padela (2007) also emphasises the concept of Islamic medical ethics in the context of *adab*, i.e., there are a set of proper conduct and professional, practical morals to be considered as a guideline of medical ethics in Islam. A comparison between the classical and modern of *adab* literature was also discussed in the paper. Although this paper clearly outlines the theory of Islamic medical ethics, there was neither an analysis nor a significant application on current real-life medical practices.

On the other hand, to provide a theoretical relevance relating to Islam under the Western concept of medical ethics, Mustafa (2014)'s paper, 'Islam and the Four Principles of Medical Ethics' investigates four key ethical principles pertinent to the Western medical field, as discussed by Beauchamp and Childress (1979), and the extent to which each principle is rooted in the Islamic literature. Thus, determining whether these principles, as a medical system or approach, are appropriate for Muslim doctors to utilise. The four principles of medical ethics are: beneficence, non-maleficence, justice, and autonomy. The author argues that "the concept of beneficence permeates through almost every layer of Islamic thought, from the material to the metaphysical" and mentions the significance of two of Allah's Names,

ar-Rahman (The Most Beneficent) and *ar-Rahīm* (The Most Merciful), several Qur'ānic verses, as well as, a *ḥadīth* (narration) to support his claim. The principle of maleficence is similar to the principle of *al ḍarar yuzāl*, that is, to remove harm, and the removal of harm takes precedence over an act with an equivalent benefit. Both Islamic and non-Islamic worlds view justice as a prerequisite when it comes to ethics, with no exception to medical ethics. As for the principle of autonomy and its relation to Islam, Islam also emphasises on an individual's responsibility and human rights' protection. However, in Islam, the Rights of Allah as the Creator are far superior to the individual rights of people, because all rights, including human rights, are determined by Allah.

Another paper that addresses medical ethics, specifically on abortion and euthanasia, from an Islamic point of view is Brockopp's (2008) 'Islam and Bioethics: Beyond Abortion and Euthanasia'. This paper discusses five articles, published in the *Journal of Religious Ethics*, relating to brain death, feeding tubes, gender selection, organ transplantation, and spiritual counselling. Brockopp (2008) also considers two broader implications, considering modern biotechnological advances: (a) the evolving definition of the human person, and (b) the ongoing importance of Islam in legitimising ethical responses to those advances. In light of the latter, there has been an attempt to establish a universal guideline on Islamic ethical principles in the medical field, thus justifying the ethical reasonings behind the general medical advances. The World Health Organisation (WHO) and the Islamic Medical Association of North America (IMANA) have each published their own guidelines on medical and health ethics from an Islamic perspective. Having discussed several available writings on Islamic medical ethics, the following sections will now discuss the three selected medical ethical issues: abortion, euthanasia, and organ transplantation. Brunei Darussalam's State Mufti's *fatāwā* pertaining to these issues will also be highlighted during this discourse.

CURRENT ETHICAL ISSUES

On Abortion

The first ethical issue relates to abortion. The burning question is: if a foetus achieves life-viability and its mother decides on abortion, then does discontinuing its life constitute murder? What about if the pregnancy is the consequence of a rape? In an Islamic juristic sense alone, abortion for any reason is considered illegal. However, as Islam emphasises on values such as compassion; it is important to reconsider the reasoning behind the abortion. Some '*ulamā*, or learned Muslim scholars, agree that in the case of abortion due to rape, abortion is only permissible before 120 days after gestation. From the Islamic perspective, the beginning of life – by Allah's *idhn*, or Permission, – usually occurs after this period. Therefore, it is forbidden to terminate a foetus after a spirit is breathed into the foetus by Allah's *idhn*. In the case of the *fatwā* issued by Brunei Darussalam's State Mufti, abortion in the case of rape is not sinful, as the action is done out of desperation (Abdul Aziz bin Juned, 2015, p. 3). Allah assures,

﴿فَمَنْ اضْطُرَّ غَيْرَ بَاغٍ وَلَا عَادٍ فَلَا إِثْمَ عَلَيْهِ إِنَّ اللَّهَ غَفُورٌ رَحِيمٌ﴾

...but if one is forced by necessity without wilful disobedience nor transgressing due limits, then there is no sin on him. Truly, Allah is Oft-Forgiving, Most Merciful. [Qur'ān, Sūrah al-Baqarah: 173]

The Prophet ﷺ also reassures that an unwilling sinful act done in a desperate situation can be forgiven by Allah. This reassurance is recorded in a *ḥadīth*, narrated by Ibn Mājah, on the authority of Ibn ‘Abbās, in which the Messenger of Allah ﷺ said, “Verily, Allah has pardoned (or been lenient with) for my *um-mah*: their mistakes, their forgetfulness, and that which they have been forced to do under duress.” (An-Nawawī). Thus, it is evident that if a rape victim aborts her baby, then she is allowed to do so, before the gestation period of 120 days, in an Islamic point of view. On the other hand, aborting an unwanted pregnancy due to voluntary adultery is forbidden at all cost, as Allah warns,

(وَلَا تَقْرُبُوا الزَّوَاجَ إِذَا كَانَ فَاحِشَةً وَسَاءَ سَبِيلًا)

And come not near to the unlawful sexual intercourse. Verily, it is a fāḥishah [(i.e., anything that transgresses its limits (a great sin)], and an evil way (that leads one to Hell unless Allah forgives him). [Qur’ān, Sūrah al-Isrā’: 32]

Unlike euthanasia and organ transplantation cases, where the Good Medical and Dental Practice guideline emphasises that doctors are to give rights to patients and families, in cases of abortion, the hospitals in Brunei Darussalam usually only deals with undeliberate abortions, as it is a crime to induce abortion, unless the doctors concluded that the mother’s health is in jeopardy (Syariah Penal Code Order 2013). In reality, there are no specific guidelines on how to handle these ethical issues, but there is a general guideline, which encourages doctors, patients and their families to make the final decisions or consent to the doctors’ suggestions. The following e-mail interviews with several hospital nurses will shed a light on how these ethical cases are actually handled in Brunei Darussalam hospitals. As for the first issue, a nursing officer, working in the Tutong Hospital, addressed the medical ethics in the cases of abortion:

In cases of abortion, patients will be admitted to the Obstetrics and Gynaecology ward through the Accident and Emergency Unit or through the Gynaecology Clinic. The patient will be monitored to determine whether it is necessary to perform the Evacuation of Retained Product of Conception (ERPC) procedure on the patient.

Another nurse, who is the Head of Obstetrics and Gynaecology Department in the Tutong Hospital, further elaborated:

[The way that the] doctor and nurses handle abortion cases depend on the case if the patient has complaints of bleeding or any complication, the doctor will do [an] ultrasound to find if any sac is still inside and to do blood test if any infection is noted. If a sac is found, [then] the doctor will admit the patient into the ward and give some antibiotic or blood transfusion if the patient’s blood Hb [level] is low. If the procedure fails, then the d&c [dilation and curettage] will [be] done in [the] operation theatre to remove the sac immediately, [supplied] continuously with antibiotic and iv [intravenous] drip. There is no such thing as complaining to upper authority. Just refer to medical social worker and ustazah [female religious teacher]. Nurses are only to follow what the doctor orders.

Having discussed Brunei Darussalam’s State Mufti’s religious opinion on abortion and two nurses’ answers on how they handle cases of abortion, the next section discusses the medical ethical issue of euthanasia.

On Euthanasia

15

At the moment, there is no exact definition of euthanasia that is agreed upon. *Stedman's Medical Dictionary* defines euthanasia as a quiet, painless death and the intentional putting to death of a person with an incurable or painful disease as an act of mercy through artificial means, whilst in the *Collins English Dictionary*, euthanasia is the deliberate act of killing someone painlessly, especially to relieve suffering from an incurable illness (Mason & McCall Smith, 1994, pp. 316). By comparing both definitions, there are common keywords consistent throughout, and they are: painless, intentional or deliberate, and incurable. Both definitions are similar in a way that to relieve suffering is deemed as an act of mercy. There are, therefore, two types of euthanasia: active and passive euthanasia.

Active euthanasia is an active intervention by a doctor to end life. In other words, an action with the intention to terminate a human life within a medical setting. Alternatively, when a doctor decides not to prolong life and withdraws treatments vital to maintaining a human life, it is passive euthanasia (*British Medical Association Report*, pp. 2–3), for example, switching off the oxygen support for patients who are unable to effectively breathe on their own. There are also other types of euthanasia, such as voluntary euthanasia, involuntary euthanasia, as well as, non-voluntary euthanasia. However, this chapter will only focus on passive euthanasia. The question now is: when an unconscious patient relies on a life support system to stay alive, who should have the final say, between the patient's family and the doctors, on switching off the life support system? For doctors, the legal indication of death is either the cessation of a heartbeat or brain death, in which a patient's heart only beats with the help of a machine (Atighetchi, 2008, p. 297). However, for Muslims, death is a Divine act when the soul separates from the body (Chishti, 1991, p. 123). If a patient relies on a life support system to stay alive, then doctors are usually certain of the patient's inexorable fate. The family of the patient may eventually learn to accept the patient's fate, or they may not.

The above issue then leads to the next question: would switching off the life support system be considered as a wise medical decision, a murder, or an act of mercy? If there is no hope in resurrecting a terminally ill patient who is in a vegetative state, then clearly, death is inevitable, and ultimately, Allah has the final decision on the patient's death. To tackle the ethical issue of euthanasia, Brunei Darussalam's State Mufti issued a *fatwā* regarding terminally ill patients who rely on life support machines. He affirmed that it is permissible to switch off the life support machine, if the patient is on the verge of death or the patient himself already gave his consent beforehand. The consideration of the patient's permission is regarded because the State Mufti judged the life support machine as a treatment, and the sick patient has the right to decide on the continuation of the treatment. However, the Islamic Fiqh Academy, an affiliate of the Muslim World League (MWL), issued a *fatwā* that prohibits euthanasia either at the request of the patient or a relative. Similarly, Omar Kasule, a prominent speaker in the field of Islamic Medical Ethics, asserts that "euthanasia violates the Purpose of the Law to preserve Life by taking life and it leads to cheapening human life making genocide more acceptable" (Omar Kasule, 2007). In the same paper, he continues to argue that euthanasia also violates the preservation of Faith as it is assumed that the cause of death is by Allah's Right, which is far from true, as there is a clear human involvement in both active and passive euthanasia.

As for cases of euthanasia, both nurses have similar answers respectively:

The doctor would meet with all of the patient's closest family members to discuss and explain the situation of the patient, and the nurse would only reinforce the information given by the doctor, and the nurse also acts an advocator for the patient himself and his family members.

The doctor will discuss with [the patient's] family members regarding terminating the use of life-supporting machine if [the] patient's prognosis is too poor, but the decision will be [made] by [the] family on whether they agreed or not.

The following section discusses the State Mufti's juristic opinion on the medical ethical issue of organ transplantation, followed by accounts of its real-life medical practice in hospitals.

On Organ Transplantation

The third and final ethical issue concerns organ transplantation, particularly on the rights of the living donor or the dead body, and the receiver. From an Islamic perspective, obtaining permission from the living donor is most important. Unless the living donor gave his permission before death, organ transplantation from the deceased to the living may not be allowed. However, the only exception for organ transplantation to be allowed in the case where the deceased did not grant his permission before death is by attaining the consent of the family of the deceased. According to Brunei Darussalam's State Mufti, if the deceased's family approves of the organ transplantation, then any organ is allowed to be transplanted to the receiver, except for organs that can be used for gender reassignment surgeries, such as ovaries. The State Mufti also emphasised several medical ethics to abide by on the process of removing organs from the deceased's body, including taking care of the corpse's *'aurah* and limiting the transplantation to only organs that are crucial for the receiver. Surgeons are also required to perform the surgery meticulously, only to the required extent, and the surgery should be performed in a quiet and peaceful environment to respect the deceased (Abdul Aziz bin Juned, 2015, pp. 114–115). On the other hand, in cases of organ transplantation, only some organ transplant procedures are performed in Brunei Darussalam hospitals. Most of the procedures are kidney transplantation, whilst other organ transplantation cases are usually sent to hospitals overseas. Both nurses agree to this fact:

[Most] cases [of organ transplantation are] done in Brunei. The transplants organ [are extracted] from [the patient's] relatives, siblings, wife or husband, with a variety of [compatibility] test done. Almost [all] kidney transplants [are] done in Brunei. [Both] patient and family will give consent for [the] organ transplantation.

For the time being, Government hospitals do not perform [some] organ transplant procedures. Patients will be sent to a hospital overseas to receive the surgical treatment or procedure of organ transplantation.

Having discussed all three medical ethical issues, the next section analyses the level of congruency between the real-life medical practices in Brunei Darussalam hospitals and the State Mufti's *fatāwā*.

SOLUTIONS AND RECOMMENDATIONS

The above interviews describe how nurses handle certain medical ethical cases, pertaining to abortion, euthanasia, and organ transplantation in Brunei Darussalam hospitals. One might ask, are these medical ethics congruent to the State Mufti's *fatāwā*? It is worth recalling that Brunei Darussalam's State Mufti declared that abortion, before the gestation period of 120 days, is not sinful, but only for rape cases. Passive euthanasia is deemed permissible, by the State Mufti, if the patient is on the verge of death, or

the patient himself already gave his consent beforehand. As for organ transplantation, the State Mufti also declared it permissible, as long as the patient granted his permission before death. Based on the nurses' answers, with the exception of organ transplantation, it is difficult to determine the circumstances of those cases, unless they are made clear. For example, in abortion cases, the patient may or may not confide to the *ustazah* (female religious teacher) the real reason behind her pregnancy, perhaps due to trauma or feelings of shame. Even for euthanasia, according to the State Mufti, the patient should be able to give his consent to switch off the life support machine. In reality, unfortunately, these cases are often decided by the patient's family instead.

Furthermore, there are no particular guidelines on certain ethical issues in the medical field, which in this case, relating to abortion, organ transplantation, as well as, euthanasia. Possibly in the future, Brunei Darussalam's Ministry of Health could consider the State Mufti's *fatāwā*, if they were to implement specific ethical guidelines for medical practitioners to apply in some complex medical cases. For the time being, clearly, there is still a level of secularisation between the medical field and religion. For example, although Brunei Darussalam's State Mufti issued a *fatwā* regarding abortion, the *fatwā* has not been included as part of the ethical guidelines in hospitals, yet. In Brunei Darussalam, medical ethics are clearly still separated from legal issues, even though in Islam, aborting a foetus is considered both an ethical, as well as, a legal issue. Perhaps, the biggest challenge that the Ministry of Health, or any other ministries, for that matter, faced is the lack of public articulation and socialisation before implementing a new strategy. The lack of socialisation and education of ideas to the public before the implementation of a new strategy may drive the public to be reluctant in accepting and welcoming the new change. If the Ministry of Health is successful in articulating new healthcare plans and guidelines that are in accordance to the *Shari'ah*, then the Bruneian society might just be more accepting of these strategies, as they understand that the implementation of these plans has not only social benefit, but also devotional significance, in terms of fulfilling the *fardhu kifāyah*.

FUTURE RESEARCH DIRECTIONS

The growing Islamic influence in Brunei Darussalam would probably lead to the implementation of more Islam-motivated health principles and medical ethics in Brunei Darussalam hospitals. But for the moment, further research is required, particularly on the functional relevance of Islam in the healthcare system. For example, when considering the existing cures recommended by Islam, mental institutions or psychiatric departments could start adopting an Islamic form of medicine, or spiritual healing, as an alternative for or in complementary with Western treatments (e.g., psychotic drugs and cognitive behavioural therapy). In this case, a different type of research is necessary to investigate whether such methods of alternative, or complementary, treatment are effective or ineffective on the mental state of psychologically ill patients. Further research could also specifically look at healthcare on the state level, on how much has the Ministry of Health incorporates Islamic principles and values into its healthcare system and organisation. This area of research would be a significant contribution to further understanding the health components of an Islamic system of governance, since people cannot function outside of a political and social organisation, i.e., no man can isolate himself from society. And creating a society conducive for the worship of Allah is the ultimate vision of an Islamic system of governance.

CONCLUSION

The findings demonstrate a certain degree of congruity, with the exception of abortion cases. Brunei Darussalam's State Mufti declared that abortion in cases of rape is not sinful, but legally, it would still be considered as a crime, as the current Brunei law (as of 2019) states that abortion is permissible only if the pregnancy is detrimental to the mother's health, as diagnosed by a medical practitioner in good faith. It is uncertain whether the Brunei Government will integrate the State Mufti's *fatāwā* into the Syariah Penal Code Order (SPCO) 2013, but on the healthcare side, absorbing the State Mufti's *fatāwā* into the medical ethical guidelines might just mitigate the disparity between healthcare and religion. If the Ministry of Health could articulate well the significance of Islamic principles and values in healthcare, then it would be another step forward for the Bruneian society to appreciate the functional relevance of Islam, perhaps not just in healthcare, but also in every socio-political aspect of daily life. If, and when, the society recognises Islam's functional relevance in its system of governance, Brunei Darussalam would then embody a society that is conducive for the worship of Allah, much like the Madīnian polity during the time of Prophet Muhammad^ﷺ. In achieving this vision, Brunei Darussalam could confidently assert to be a 'Negara Zikir', or 'Zikr Nation', as envisioned by the Sultan.

REFERENCES

- Abdul Aziz bin Juned. (2015). *Rawatan dan Kesihatan 2: Himpunan Fatwa Mufti Kerjaan Negara Brunei Darussalam Mengenai Amalan, Rawatan dan Kesihatan 2002-2013*. Brunei Darussalam: Jabatan Mufti Kerajaan, Jabatan Perdana Menteri.
- Al-Bar, M. A., & Chamsi-Pasha, H. (2015). Assisted Reproductive Technology: Islamic Perspective. In *Contemporary Bioethics: Islamic Perspective*. Springer. doi:10.1007/978-3-319-18428-9_11
- An-Nawawī. (n.d.). *Commentary of the Forty Ḥadīth of al-Nawawī Vol. 2* (J. al-Din M. Zarabozo, Trans.). Al-Basheer Publications & Translations.
- Atighetchi, D. (2009). *Islamic Bioethics: Problems and Perspectives* (1st ed.). Springer.
- Beauchamp, T. L., & Childress, J. F. (1979). *Principles of Biomedical Ethics* (1st ed.). Oxford University Press.
- Beauchamp, T. L., & Childress, J. F. (2013). *Principles of Biomedical Ethics* (7th ed.). Oxford University Press.
- Brockopp, J. E. (2008). Islam and Bioethics: Beyond Abortion and Euthanasia. *The Journal of Religious Ethics*, 36(1), 3–12. doi:10.1111/j.1467-9795.2008.00333.x
- Chishti, H. M. (1991). *The Book of Sufi Healing*. Inner Traditions International.
- Council of the World Muslim League. (1990). *Proceedings of Conference of The Fiqh (Islamic Jurisprudence)*. Author.
- Elisseeff, V. (2000). A Brunei Sultan of the Early Fourteenth Century — A Study of an Arabic Grave-stone. In *The Silk Roads: Highways of Culture and Commerce*. Berghahn Books.
- Euthanasia*. (n.d.). British Medical Association Report.

Gatrad, A. R., & Sheikh, A. (2001). Medical Ethics and Islam: Principles and Practice. *Archives of Disease in Childhood*, 84(1), 72–75. doi:10.1136/ad.84.1.72 PMID:11124793

Kasule, O. (2007). Islamic Medical Ethics with Special Reference to Maqāṣid al-Sharī'ah. *I-Epistemology*. Accessed on the 18th of August 2018. Retrieved from: <https://i-epistemology.net/subjects/medicine/item/365-islamic-medical-ethics-with-special-reference-to-maqasid-al-shari-at.html>

Khan & al-Hilali. (1996a). Interpretation of the Meanings of the Noble Qur'ān in the English Language. *Sūrah al-Baqarah*, 173.

Khan & al-Hilali. (1996b). Interpretation of the Meanings of the Noble Qur'ān in the English Language. *Sūrah al-Isrā'*, 32.

Mason & McCall Smith. (1994). *Law and Medical Ethics* (4th ed.). Butterworths.

Mustafa, Y. (2014). Islam and the Four Principles of Medical Ethics. *Journal of Medical Ethics*, 40(7), 479–483. doi:10.1136/medethics-2012-101309 PMID:23975951

Padela, A. I. (2007). Islamic Medical Ethics: A Primer. *Bioethics*, 21(3), 169–178. doi:10.1111/j.1467-8519.2007.00540.x PMID:17845488

Shomali, M. A. (2008). Islamic Bioethics: A General Scheme. *Journal of Medical Ethics and History of Medicine*. Retrieved from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3713653/>

Syariah P. C. O. (2013). Retrieved from: <http://srhr.org/abortion-policies/documents/countries/01-Brunei-Darussalam-Syariah-Penal-Code-Order-2013.pdf>

The Good Medical and Dental Practice. (n.d.). Retrieved from: <http://www.moh.gov.bn/Shared%20Documents/Brunei%20MEDICAL%20BOARD/Good%20Medical%20and%20Dental%20Practice.pdf>

The Oxford Dictionary. (n.d.). Retrieved from: <https://en.oxforddictionaries.com/definition/ethics>

Union of News Agencies of the Member States of the OIC. (2015). *Islamic Fiqh Academy Issues Fatwa Prohibiting Euthanasia*. Accessed on the 25th of March 2018. Retrieved from: http://iinanews.org/page/public/news_details.aspx?id=87090&NL=True#.WuebIYhubIU

World Medical Association Declaration on the Right of the Patient. (1981). Accessed on the 26th of August 2017. Retrieved from: <http://dl.med.or.jp/dl-med/wma/lisbon2005e.pdf>

ADDITIONAL READING

Ahmed, A. A. Y. (2013). *The Fiqh of Medicine: Responses in Islamic Jurisprudence to Developments in Medical Science*. Ta-Ha Publishers.

Chapra, M. U. (2008). *The Islamic Vision of Development in the Light of Maqāṣid al-Sharī'ah*. The International Institute of Islamic Thought.

Islamic Medical Association of North America Ethics Committee. *Islamic Medical Ethics: The IMANA Perspective*. Retrieved from: <http://fimaweb.net/main/medicalethics/imanaethicspaper.pdf>

Kamali, M. H. (1991). *Principles in Islamic Jurisprudence*. The Islamic Texts Society.

Saifuddeen, S. M., Chang, L. W., Ibrahim, A. H., & Nor, A. M. K. (2013). Islamic Ethical Framework to Tackle Scientific and Technological Dilemmas. *Journal of Dharma: Journal of Religions and Philosophies*, 38(4), 373–386. https://umexpert.um.edu.my/file/publication/00010699_159305_70842.pdf

Saifuddeen, S. M., Rahman, N. N. A., Isa, N. M. and Baharuddin, A. (2013). Maqāṣid al-Sharīʿah as a Complementary Framework to Conventional Bioethics. *Journal of Science and Engineering Ethics*.

Siddiqui, A. (1997). Ethics in Islam: Key Concepts and Contemporary Challenges. *Journal of Moral Education*, 26(4), 423–431. doi:10.1080/0305724970260403 PMID:11656902

World Health Organisation. Regional Office for the Eastern Mediterranean. (2005). *Islamic Code of Medical and Health Ethics*. Fifty-second Session, Agenda item 8. Retrieved from: https://applications.emro.who.int/docs/EM_RC52_7_en.pdf

KEY TERMS AND DEFINITIONS

Abortion: An intentional act of eliminating an embryo or foetus, thus terminating the pregnancy.

‘Aurah: Parts of a body required to be covered by clothing in the presence of a non-*maḥram* (a person to whom marriage is permissible).

Euthanasia: An act of ending a patient’s life, either by injecting lethal solutions or depriving the patient of necessary treatments and/or nutrients.

Farḍhu kifāyah: Obligations to reach social sufficiency in an Islamic society. Failure to fulfil these social obligations would place the society under a huge sin.

Fatāwā (sing. Fatwā): Religious rulings that pertain to matters of the *Sharīʿah*.

Ijtihād: The reasonings, or opinions, or interpretations of a Muslim based on his or her understanding of the *Sharīʿah*.

Islamic Medical Ethics: When Islamic principles are applied as a guidance for ethical or moral behaviour in the medical and scientific fields that especially involve human life.

Organ Transplantation: A surgical procedure of removing a deceased patient’s organ and that organ is placed in the living beneficiary’s body as a replacement for the beneficiary’s damaged or missing organ.

Shahādah: Translates to “testimony” or “witness” in English. It is the first of the five pillars of Islam and is divided into two parts: *ashadu ‘alla ilāha illal’Lāh* (I testify that there is no god except Allah), and *wa ashadu ‘anna Muḥammadan rasūlul-Lāh* (and I testify that Muhammad^ﷺ is the Messenger of Allah).

Sharīʿah: Translates to “the way” in English. It is the Islamic Law that is derived from the Qur’ān and the Sunnah (Prophet Muhammad^ﷺ’s actions, sayings, and tacit approval).

APPENDIX**15**

The e-mail interview questions, pertaining to cases of abortion, euthanasia, and organ transplantation, are as follows:

- 1) How do doctors and nurses handle abortion cases in hospitals? Or are abortion cases being handled in private clinics instead?
- 2) How do doctors and nurses handle cases of euthanasia (a decision not to prolong life by terminating the use of life-supporting machines)? Who makes the decision to remove the life support? The medical team, the family or the patient?
- 3) How do doctors and nurses handle organ transplant cases? What kind of organ transplants are being performed in the hospital (e.g. heart, kidney, liver, etc)? Who gives consent for organ transplants, the patient or the family?

Distribution of Selected Health Technology in Regions of Slovakia

Beata Gavurova

Technical University of Košice, Slovakia

Matus Kubak

Technical University of Kosice, Slovakia

INTRODUCTION

Demographic aging processes in the countries have a major impact on the financial sustainability of healthcare systems. In Slovakia, as in many other countries, the proportion of the economically inactive population is increasing. The financial sustainability of the health system is very sensitive not only to aging society, but also to economic shocks. Demographic shift - increasing the share of the economically inactive population causes wage-dependent premiums to fall, while health spending may increase due to higher healthcare costs for older citizens. Therefore, the resilience and sustainability of the health system need to be ensured on an ongoing basis and prepared for changes in demographic structures in the countries. Attention is drawn to the role of ensuring more efficient use of available resources, and also maximizing health and well-being of the population (Dlouhý, 2009; Fragkiadakis et al., 2016). The health system's performance consists of three components: efficiency, availability and resilience. These three economic categories are also influenced by the distribution of medical devices and equipment, both by the quality of satisfaction of medical needs, by geographical availability and by the cost - the degree of financial burden on the health system. However, only few research studies deal with the issue of regional disparities in the use of selected medical technology. It is especially caused by problematic access to the data base, as well as insufficient attention to examining the impact of prevention programs on the effectiveness of the health system and their connectivity to early diagnosis and treatment. These facts present the main cause for executing this study. It aims in investigating regional disparities in the use of magnetic resonance imaging (MRI) in Slovakia during the years of 2008 – 2017, and their importance for assessing the availability and quality of healthcare in the Slovak health system.

BACKGROUND

Many foreign research studies examine the use and distribution of health technology in the regions of their country to explore different aspects of health care systems. Some research studies investigate health technology as a subject of a research, other than as a part of the methods in which they act as variables. As early as 1988, Stephan (1988) conducted a number of technology assessments in his study to help policy makers in decision making. The author states that despite the efforts, a certain degree of uncertainty in the assessment of health technology will always persist, so decision-makers must also accept the level of arbitrariness. Many authors have examined the distribution of health technology to ensure equal access to healthcare and eliminate regional disparities in their studies. Xing and Dyama (2016)

DOI: 10.4018/978-1-7998-3479-3.ch130

explored the burden of local public hospitals and their impact on the health care system. They identify a number of issues related to higher quality of healthcare and also the connection to technological innovation. Winchester and King (2018) analyse an uneven approach to healthcare in their study, declaring that decentralization cannot resolve this uneven approach in the country. Seymour et al. (2006) examined geographical aspects of imaging diagnostic equipment (IDE) for stroke suspected patients in Scotland. They used geographic information systems and analysed the availability of these health services. The authors point to the limited availability of stroke scanning services during weekends in the individual regions, and identify a number of recommendations for health policy makers. The macroeconomic view of this issue was the result of a study by Suzuki et al. (2018). Their study examined and analysed the relationship between the economic conditions of the OECD member countries and available resources in the health systems. The authors primarily focused on exploring the factors of implementation and use of the IDE. Altogether, 29 countries were classified into four groups according to medical environment and economic indicators. The number of hospital beds has been shown to have the greatest impact on CT and MRI numbers. The authors also plan to clarify other factors influencing the implementation and use of the IDEs by investigating high-growth medical equipment market. One of the causes of regional disparities in implementing of new IDEs may be an increasing complexity of health technologies. Also, it is proved by the results of Liljegren et al. (2006) study. The authors analyse the reasons of all the obstacles in implementing complex IDEs and look for appropriate methods to monitor these reasons. Marinic and Varga (2016) analysed diagnostic procedures and equipment for healthcare facilities in the context of the use of human resources in specialized fields: anesthesiology, gastroenterology, neurosurgery, pathology, microbiology, electrophysiology, neurology and cardiology. Data for analysis were obtained from all hospitals in Croatia in 2013 and the results were compared with available international recommendations. The results of the analyses pointed to significant differences between Croatian hospitals as well as international standards in Cost Utilization. The study has brought many implications to health policy. The research of organisational impact in the field of medical devices was researched by Roussel et al. (2016). Health Technology Assessment (HTA) is a rapidly developing area. Non-clinical areas have also been assessed in recent years. The authors emphasize the health-economic aspect, which has been the most prominent in the assessment of health technologies, while the organisational impact assessment has hardly been explored. Shah et al. (2015) investigated in their study an availability of health technology in particular types of healthcare - in injuries in Nepal. Their aim was to identify the factors causing regional disparities in the availability of health technology. The authors propose solutions to be cost effective and sustainable, linking them to an optimal planning process. The methodological aspect in relation to the use of health technology is examined in the study of, e.g. de Moraes et al. (2010). The authors investigated whether the Multicriteria Decision Aid methodology supports the creation of performance benchmarks within health technology processes. Also, the authors emphasize the need for multidisciplinary cooperation in the field of healthcare as well as cooperation between the different levels of management: operational and strategic. As authors state, the assessment of health technologies may only be successful with a multidisciplinary team approach. Similarly, the economic categories frequently appear in the studies as the target aspects of the IDE's research and use besides the technological processes. For instance, Wei et al. (2018) studied the relationship between hospital efficacy, and CT and Magnetic Resonance Imaging (MRI) in their study. They selected a random sample of 131 hospitals in China and conducted a questionnaire survey. Hospital efficiencies were measured by SFA, and subsequently planned to use DEA. A sample of hospitals was not chosen equally from a geographical perspective. The study also did not analyse inappropriate and appropriate use of CT and MRI, which the authors also report as a limitation of their study. Many authors used medical equipment

and devices as input variables within their analyses. Grosskopf and Valdmanis (1987), Zuckerman et al. (1994), Puig-Junoy (1998) and others used medical devices as a whole, as a variable. Bradford et al. (2001) used as a variable the usage of established and new technologies, Dey et al. (2006) used a capital equipment as a variable, which was defined as a current technology equipment to measure performance. Henry et al. (2013) investigated the impact of using new health technologies for early diagnosis and treatment, Tsekouras et al. (2010) monitored the impact of implementing new high-tech health technologies to increase production efficiency, Kounetas and Papathanassopoulos (2013) observed changes in the efficiency of Greek hospitals by introducing state-of-the-art medical devices. Many authors, such as Garmer et al. (2004), Hounbo et al. (2017), Postiche et al. (2018) had examined only selected types of medical devices, e.g. CT and MRI. Studies by Samut and Cafri (2015) belong to more recent studies and they are also interesting as they use as input variables number of MRI and CT devices calculated per 100.000 inhabitants, or Lee (2016) who used as one of the input variables the number of the MRI units, and others. The studies of Gavurova et al. (2017, 2018) were conducted in the Slovak environment and analysed the distribution of medical equipment and devices during 2008-2014. The presented study is linked to these results in order to investigate the regional disparities in MRI distribution in Slovakia, while using the correspondence analysis. The selection of this graphical method was determined by the objective of the study and the potential implications for health policy makers. The results of the analyses will be beneficial in the processes of creating national and regional strategic health plans, as well as in developing a strategic framework in the health system with an emphasis on assessing the availability of healthcare. This availability is linked not only to prevention programs as an input component of the health system efficiency, but also to the overall quality of health care as an output component, measured by the level of morbidity and mortality that is highly differentiated in individual regions of Slovakia.

DATA AND METHODOLOGY

In this study, we are focusing on the distribution of the magnetic resonance imaging scans within the timespan 2008-2017 in district town in Slovak republic. Dataset used in this analyse was provided by National Health Information Centre (NHIC), the state-funded organization founded by the Ministry of Health of the Slovak Republic. Slovak republic is divided on 79 districts (NUTS 4 level), among which 10 belongs to a two biggest cities Bratislava, and Kosice. And therefore, we distinguish 71 cities. Figure 1 depicts total magnetic resonance imaging scans in Slovak republic within time period 2008-2017. We can see on the figure that during period of 2008-2017, the total number of magnetic resonance imaging scans has increased from 33 magnetic resonance imaging scans in 2008 to 51 scans in 2017.

Figure 1. Total number of magnetic resonance imaging scans in Slovak republic within 2008-2017

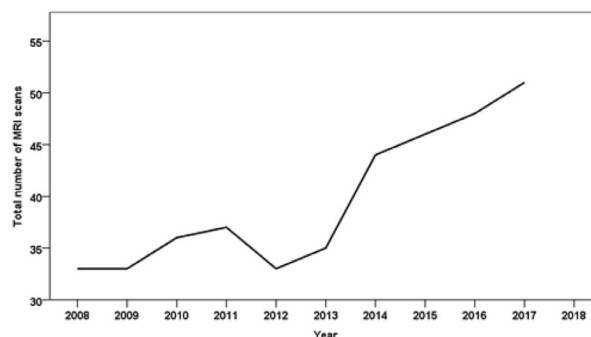
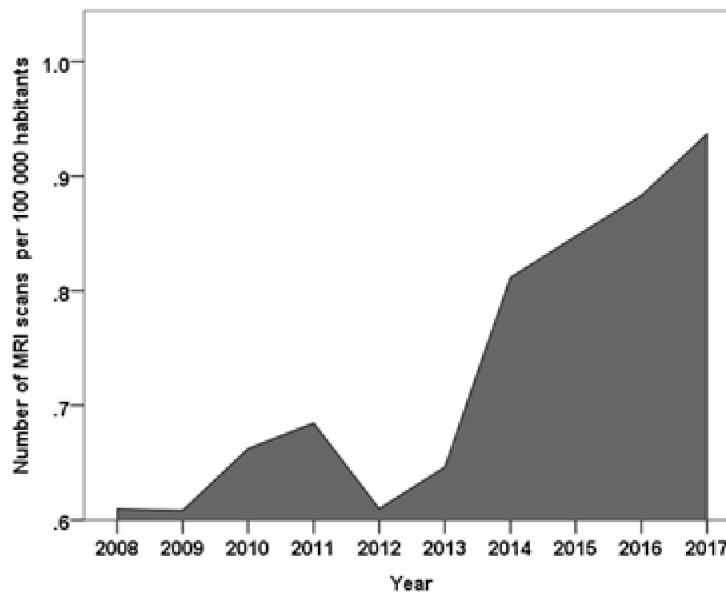


Figure 2 describe the distribution of magnetic resonance imaging scans per 100 000 habitants during period of 2008-2017 (by the 31st of December 2018, Slovak republic had 5 450 421 of habitants). One can see, that magnetic resonance imaging scans has increased over time approximately 0.6 scans per 100 000 habitants in 2008 to approximately 0.94 scans per 100 000 habitants in 2017.

Figure 2. Number of magnetic resonance imaging scans per 100 000 habitants within 2008-2017



The distribution of magnetic resonance imaging scans over time among regions on NUTS 3 level in Slovak republic is shown on figure 3. Slovak republic is divided on the 8 regions on NUTS 3 level. In this study we distinguish 8 regions on NUTS 3 level, but we also isolated two biggest cities, Bratislava city and Kosice city, which have a large catchment area. Figure 3 indicates, that distribution of magnetic resonance imaging scans is profusely concentrated in favour of two big cities, Bratislava city and Kosice city. Bratislava city is strongly outlying even if compared to Kosice city. We see on figure 3, evident increase in number of magnetic resonance imaging scans per 100 000 habitants in Bratislava city. In Kosice city we document decrease in term of number of magnetic resonance imaging scans per 100 000 habitants over time. In other regions, where the number of magnetic resonance imaging scans per 100 000 habitants is close to 1 we document on average slight increase of number of scans over time.

Figures 4-11 describe magnetic resonance imaging scans equipment per 100 000 habitants in 8 regions in Slovak republic on NUTS 4 level during observed period. Figure 4 depicts situation in Bratislava region. We can see that best equipped is the Bratislava city within its city districts. All other cities have no MRI scans, except Malacky city from 2015.

On Figure 5, we can observe the number of MRI scans per 100 000 habitants in Trnava region during observed period. As in case of other region seat cities, we can observe the good equipment in Trnava city. From 2012, we observe very good facilities in Skalica city. By the end of the observed period we observe amelioration of situation in Dunajská Streda and Piestany.

Figure 3. Distribution of magnetic resonance imaging scans among regions in Slovak republic

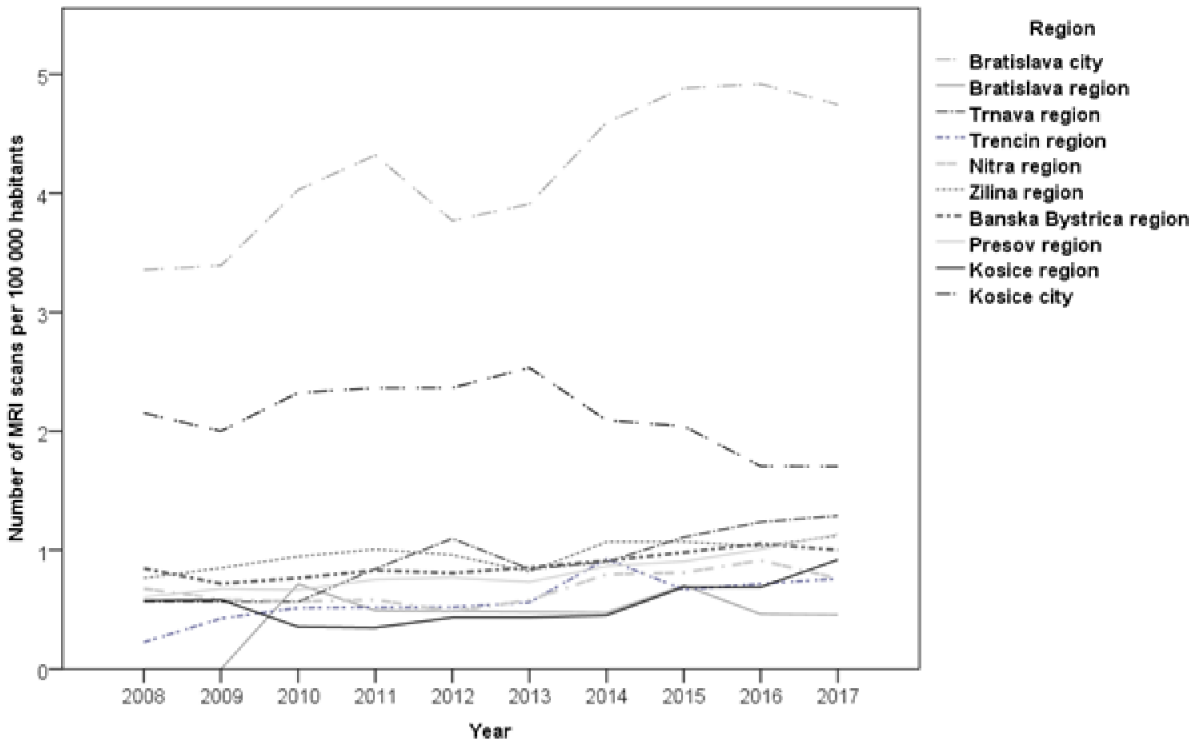


Figure 4. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Bratislava region

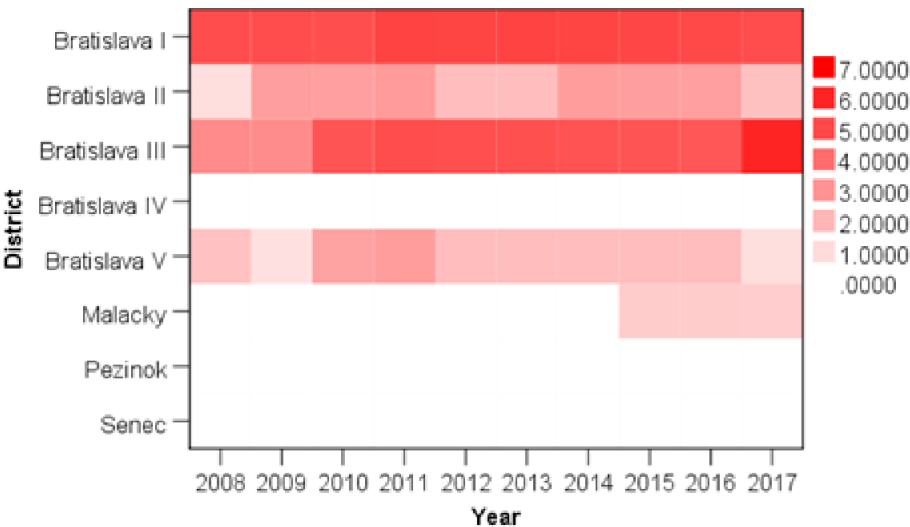


Figure 5. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Trnava region



The situation in Trenčín region is on the Figure 6. We can conclude, that the region as a whole is underequipped by MRI scans and even the region seat city is well equipped only from 2014. Another middling equipped city in Trenčín region is Prievidza city.

Figure 6. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Trenčín region

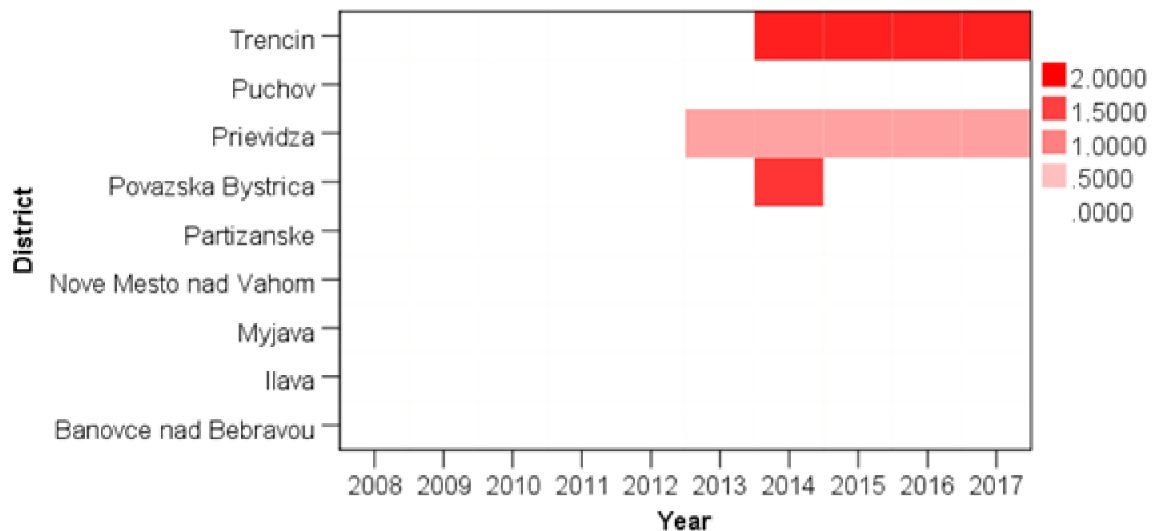
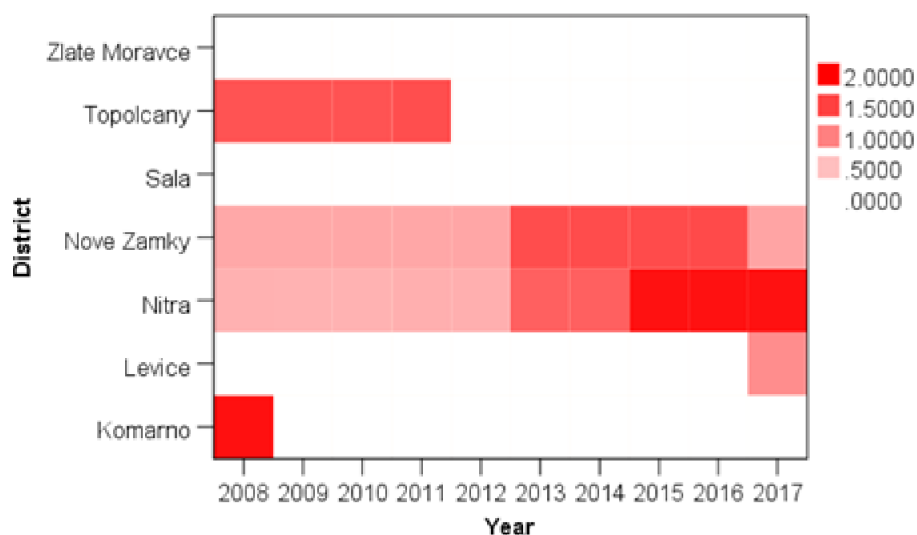


Figure 7 captures situation in Nitra region. Here we can note that by the beginning of observed period, the region was spatially well equipped by MRI scans. By the end of the observed period, three cities are well equipped by MRI scans. Nitra and Nove Zámky are well equipped in long run.

Distribution of Selected Health Technology in Regions of Slovakia

Figure 7. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Nitra region



Situation in Zilina region is captured on Figure 8. In Zilina region we identify two cities which are well equipped by MRI scans from long run view, Zilina and Martin. Ruzomberok and Dolny Kubin are well equipped by MRI scans from short run view.

Figure 8. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Zilina region



Figure 9 review the situation in Banská Bystrica region. Here we observe the amelioration of situation in Banská Bystrica region over time. From long run view, also Rimavská Sobota is well equipped city. From short run, Lucenec city is averagely equipped city.

Figure 9. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Banská Bystrica region



Figure 10 captures situation in Presov region, which is the most populated region in Slovakia. Here we identify Poprad and Presov as well-equipped cities from long run view. On the other hand, we can see that Stara Lubovna and Humenne are well equipped cities in short time view. Levoca city starts to be very well equipped city in 2017.

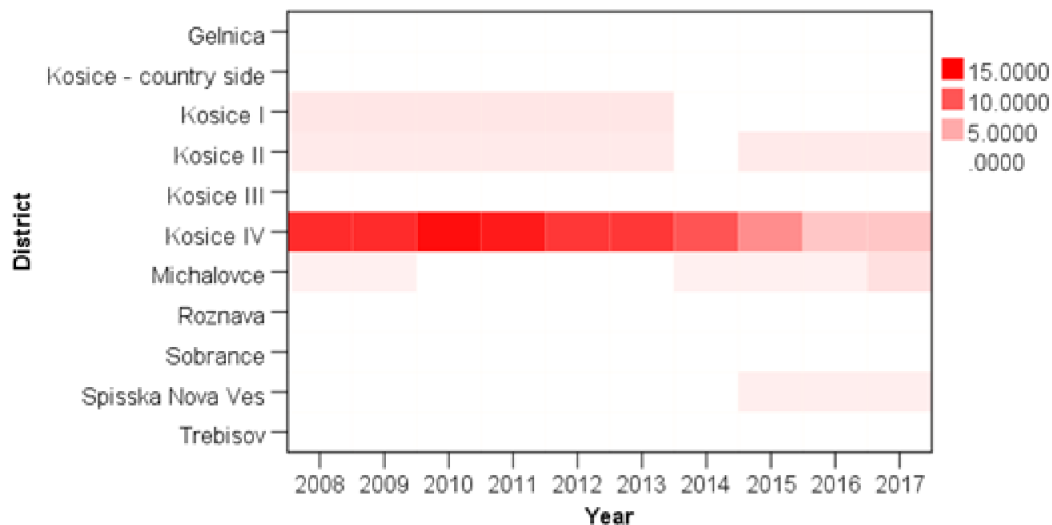
Figure 10. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Presov region



Situation in Kosice region, the second most populated region in Slovakia can be seen on figure 11. We can see, that in long run view, Kosice IV is the best equipped city district, but situation within this

area has degraded over time to standard observed in other cities in Slovakia. Otherwise, Kosice I and Kosice II are medium equipped areas. From short time view, Michalovce and Spisska Nova Ves start to be mediocre equipped cities.

Figure 11. Distribution of magnetic resonance imaging scans per 100 000 habitants among districts in Kosice region



Correspondence Analysis

Correspondence analysis is an exploratory quantitative technique intended for analysing data in pivot tables which aim is to reduce dimensionality. Correspondence analysis provides certain measures of correspondence between columns and rows of pivot table. Correspondence analysis is a popular technique especially when processing large pivot tables, where the graphical representation of the result is clearer than numeric outputs. This technique is descriptive and exploratory, which does not include testing of the statistical significance of the acquired results. Correspondence analysis is multidimensional visual technic that enable to graphically display columns and rows of the columns of the pivot table and specify the measure of correspondence between the columns and rows. The provided results are similar to results of factor analysis. Correspondence analysis deals with similar issues as the principal component analysis. In the correspondence analysis we study the relationships between the individual categories of two categorical variables. The output of correspondence analysis is two dimensional correspondence map. Coordinates in graph are analogous to factors in a principal component analysis. Correspondence analysis uses nominal variables as inputs and necessitate an existence of the dependence between the observed data.

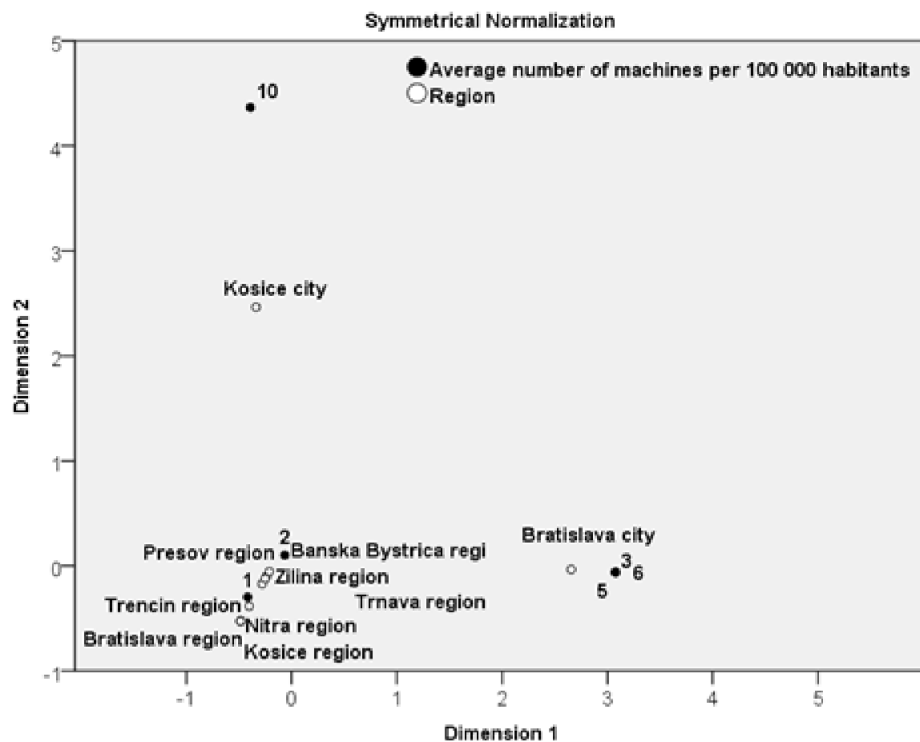
In our study, we focus to determine similarities among district cities in Slovakia, according to average numbers of magnetic resonance imaging scans per 100 000 habitants during period of 2008-2017. To process data, we use a correspondence analysis. The quality of transformation of multidimensional points into the correspondence map is assessed by total inertia. The information about total inertia of our model is in Table 1. Our model captures by its two dimensions 83.1% of total inertia. First dimension contributes by 58.2% of inertia and second one contributes by 24.9% of inertia. It can be concluded, that our model captures 83.1% of reality, what is according to nature of processed data sufficient score.

Table 1. Proportion of inertia

Dimension	Singular Value	Inertia	Proportion of Inertia		Confidence Singular Value	
			Accounted for	Cumulative	Standard Deviation	Correlation
						2
1	.863	.745	.582	.582	.116	.006
2	.564	.319	.249	.831	.223	
3	.465	.216	.169	1.000		
Total		1.279	1.000	1.000		

Figure 12 is the correspondence map of analysis of number of magnetic resonance imaging scans in district cities in Slovak republic. Here we use the average number of magnetic resonance imaging scans per year during the observed period 2008-2017.

Figure 12. Correspondence analysis: Region/Number of magnetic resonance imaging scans per 100 000 habitants



The first finding is, that Kosice city and Bratislava city form outlying entities, where Kosice city corresponds to 10 magnetic resonance imaging scans per 100 000 habitants and Bratislava city to 3, 5 and 6 magnetic resonance imaging scans per 100 000 habitants. Cities in Presov region, Banska Bystrica region, Zilina region and Trnava region form second cluster that correspond to 2 magnetic resonance

imaging scans per 100 000 inhabitants. Finally, Kosice region, Nitra region, Trencin and Bratislava region forms third cluster that corresponds to 1 magnetic resonance imaging scan per 100 000 inhabitants.

FUTURE RESEARCH DIRECTIONS

In the Slovak health system, a process of hospitals' stratification is being prepared that should ensure its higher quality and efficiency. A major modernization of inpatient health care is planned as a part of the stratification of hospitals, while both, its efficiency and also improvement of the quality of health care are important. These parameters are linked to both, availability and health care security. All those parameters are influenced by structure and use of health technology in both, direct or indirect interactions.

Demographic development is also related to changes in health needs in Slovakia, which is also reflected in the number of hospitalizations. In 2017, there were registered 1.017 mil. of hospitalizations, while 72% of them were registered for acute inpatient health care. Consequently, up to 32 000 beds were used to treat them, while inpatient health care was provided by 139 general and specialized hospitals. There are also changes evident in the hospitalization time as the patient in Slovakia spends more time in the hospital than patients in the countries of the Western Europe with regard to the international comparison. Even if there is a higher rate of hospitalization in Slovakia, there is a lower rate of bedding.

Health policy experts are extensively addressing the sustainability of health and social systems based primarily on economic and demographic developments in Slovakia, as well as on changes in morbidity and health care demand, depending on the type of morbidity. Innovations and trends in healthcare are equally important as they affect the length and quality of treatment, as well as timely and quality diagnosis. By 2030, the hospitalization rate is expected to increase by up to 18%. These consequent facts encourage to perform a number of follow-up researches to investigate the impact of health care technology on morbidity and mortality, while taking into account the economic and geographical availability of healthcare in Slovakia. The projected increase in hospitalization rates is also associated with the planned change in morbidity patterns in the elderly as well as in prolongation of life expectancy. Early diagnosis of chronic diseases, which is reflected not only in economic aspects, but also in the quality of human life, is also drawn to attention. The subsequent research goal is to investigate the impact of availability of diagnostic and therapeutic medical equipment on the change in mortality and morbidity structure of the population. The extent to which the distribution of diagnostic and therapeutic medical equipment has an impact on the development of hospitalization as well as on changes in non-communicable diseases (NCD) is interesting. Similarly, those factors that determine the imbalance in the distribution of diagnostic and therapeutic techniques and reveal the barriers to their better use need to be examined in a cooperation with the health system institutions. This will serve as a motivation in creating optimal regulation mechanisms in order to balance healthcare availability and to eliminate regional disparities in a health. These research goals may also encompass the issue of prevention programs, especially in those areas that suffer from a negative development of some diseases for a longer period of time.

CONCLUSION

The growth of health care expenditures in Slovakia is considered the second highest of all the EU Member States, as confirmed by analytical outputs from the Ministry of Health of the Slovak Republic. Slovakia is one of the few countries that did not reduce the overall health care budget during the financial crisis.

In hospital and primary health care, there are persistent issues that need to be addressed in the context of demographic aging processes as well as the sustainability of the health system. The latest Strategic Framework on Health for 2014-2030 has made clear efforts to improve cost-effectiveness. The health of the population affects the country's economic growth and is influenced by many factors, and with increasing age, co-morbidities, complicating and increasing the cost of treatment, are also increasing. Strong technological advances bring many innovations in diagnosis and treatment, while new medical practices and the use of the latest technology place higher financial demands on the country's health system. Recently, the impact of medical technology on multiple aspects of the health system has been increasingly exploring and revealing ways to quantify and economize it. So far, such studies have been absent in Slovakia. The reason is a problematic data base as well as methodological difficulty. Therefore, the aim of this study was to investigate regional disparities in the use of MRI in Slovakia for the period of 2008 – 2017, and their importance in assessing the availability and quality of health care in the health system of Slovakia. As the analysis shows, the distribution of MRIs in Slovakia is intensively concentrated in large cities Bratislava and Kosice. The city of Bratislava has significantly more MRIs in comparison to the city of Kosice. In Kosice, there is evident decrease in the number of MRI per 100.000 inhabitants over the period under review. In other regions, the number of MRI per 100.000 inhabitants was close to 1, while on average, there is an increase in MRI per 100.000 inhabitants during 2008-2017. In Kosice, there are 10 MRIs per 100.000 inhabitants on average, and in Bratislava and its districts, there are 3, 5 and 6 MRIs per 100.000 inhabitants. The second cluster consists of the cities in Presov Region, Banska Bystrica Region, Zilina Region and Trnava Region, where there are 2 MRIs per 100.000 inhabitants. The third cluster with 1 MRI per 100.000 inhabitants consists of: Kosice Region, Nitra Region, Trencin Region and Bratislava Region. These results show considerable regional disparities in the availability of MRI in individual regions of Slovakia. These facts need to be studied in the context of the region's demographic structure, epidemiological analysis, as well as the number and types of healthcare facilities. The issue of optimum availability to specialized medical procedures requiring the use of MRI, resp. others is also raised. Insufficient availability may result in late diagnosis of very serious diseases that are highly costly to treat, accompanied by deteriorating quality of life. The research of MRI use in diagnostic and therapeutic processes for individual diagnoses and their linkage to the economic base remains research focus of this research team. Another important research area may be the investigation of the process and economic usability of MRI in the individual locations and a creation of potential regulatory mechanisms to eliminate regional differences. These may also affect the elimination of regional disparities in a health of the population, which is also declared in several international health programs. The results of these analyses will be beneficial for regional health plan makers as well as health policy professionals. The results of the study will support the implementation of comparative analyses and the development of national and international health benchmarks.

ACKNOWLEDGMENT

This research was funded by the Slovak Research and Development Agency under the contract No. APVV-17-0360: “Multidimensional analysis of significant determinants of public procurement efficiency with emphasis on the application of Health Technology Assessment in the procurement preparation phase”.

REFERENCES

- de Moraes, L., Garcia, R., Ensslin, L., da Conceição, M. J., & de Carvalho, S. M. (2010). The multicriteria analysis for construction of benchmarks to support the Clinical Engineering in the Healthcare Technology Management. *European Journal of Operational Research*, 200(2), 607–615. doi:10.1016/j.ejor.2009.01.018
- Dey, P. K., Hariharan, S., & Clegg, B. T. (2006). Measuring the operational performance of intensive care units using the analytic hierarchy process approach. *International Journal of Operations & Production Management*, 26(8), 849–865. doi:10.1108/01443570610678639
- Dlouhý, M. (2009). Efficiency and Productivity Analysis in Health Services. *Prague Economic Papers*, 18(2), 176–184. doi:10.18267/j.pep.348
- Fragkiadakis, G., Doumpos, M., Zopounidis, C., & Germain, C. (2016). Operational and economic efficiency analysis of public hospitals in Greece. *Annals of Operations Research*, 247(2), 787–806. doi:10.1007/s10479-014-1710-7
- Garmer, K., Ylven, J., & Karlsson, I. M. (2004). User participation in requirements elicitation comparing focus group interviews and usability tests for eliciting usability requirements for medical equipment: A case study. *International Journal of Industrial Ergonomics*, 33(2), 85–98. doi:10.1016/j.ergon.2003.07.005
- Gavurova, B., Balloni, J. A., Tarhanicova, M., & Kovac, V. (2018). Information and communication technology in the role of information system of healthcare facility in the Slovak Republic. *Economies*, 3(3), 2227–7099. doi:10.3390/economies6030047
- Gavurova, B., Kovac, V., & Fedacko, J. (2017). Regional disparities in medical equipment distribution in the Slovak Republic – a platform for a health policy regulatory mechanism. *Health Economics Review*, 7(39), 1–13. doi:10.1186/s13561-017-0176-0 PubMed
- Grosskopf, S., & Valdmanis, V. (1987). Measuring hospital performance: A nonparametric approach. *Journal of Health Economics*, 6(1), 89–107. doi:10.1016/0167-6296(87)90001-4 PubMed
- Henry, K. A., Sherman, R., Farber, S., Cockburn, M., Goldberg, D. W., & Stroup, A. M. (2013). The joint effects of census tract poverty and geographic access on late-stage breast cancer diagnosis in 10 US States. *Health & Place*, 21, 110–121. doi:10.1016/j.healthplace.2013.01.007 PubMed
- Houngbo, P. T., Zweekhorst, M., Bunders, J., Coleman, H. L. S., Medenou, D., Dakpanon, L., & Buning, T. D. C. (2017). The root causes of ineffective and inefficient healthcare technology management in Benin public health sector. *Health Policy and Technology*, 6(4), 446–456. doi:10.1016/j.hlpt.2017.06.004
- Korolija Marinic, D., & Varga, S. (2016). Cost Utility Analysis Of Selected Medical Equipment And Procedures In Croatia. *Clinical Therapeutics*, 38(10, Supplement), e15. doi:10.1016/j.clinthera.2016.07.121 PubMed
- Kounetas, K., & Papathanassopoulos, F. (2013). How efficient are Greek hospitals? A case study using a double bootstrap DEA approach. *The European Journal of Health Economics*, 14(6), 979–994. doi:10.1007/s10198-012-0446-z PubMed

Lee, K. S., Chun, K. H., & Lee, J. S. (2008). Reforming the hospital service structure to improve efficiency: Urban hospital specialization. *Health Policy (Amsterdam)*, 87(1), 41–49. doi:10.1016/j.healthpol.2007.10.003 PubMed

Liljegren, E. (2006). Usability in a medical technology context assessment of methods for usability evaluation of medical equipment. *International Journal of Industrial Ergonomics*, 36(4), 345–352. doi:10.1016/j.ergon.2005.10.004

Postiche, H., Eghtesad, S., Kamangar, F., Etemadi, A., Keshtkar, A. A., Hekmatdoost, A., Mohammadi, Z., Mahmoudi, Z., Shayanrad, A., Roozafzai, F., Sheikh, M., Jalaeikhoo, A., Somi, M. H., Mansour-Ghanaei, F., Najafi, F., Bahramali, E., Mehrparvar, A., Ansari-Moghaddam, A., Enayati, A. A., ... Malekzadeh, R. (2018). Prospective Epidemiological Research Studies in Iran (the PERSIAN Cohort Study): Rationale, Objectives, and Design. *American Journal of Epidemiology*, 187(4), 647–655. doi:10.1093/aje/kwx314 PubMed

Puig-Junoy, J. (1998). Technical efficiency in the clinical management of critically ill patients. *Health Economics*, 7(3), 263–277. doi:10.1002/(SICI)1099-1050(199805)7:3<263::AID-HEC320>3.0.CO;2-I PubMed

Roussel, C., & Carbonneil, C., & Audry. (2016). Organisational impact: Definition and assessment methods for medical devices. *Therapie*, 71(1), 83–96. doi:10.1016/j.therap.2016.01.003

Samut, P. K., & Cafrı, R. (2015). Analysis of the Efficiency Determinants of Health Systems in OECD Countries by DEA and Panel Tobit. *Social Indicators Research*, ●●●, 1–20.

Seymour, J., Cairns, J., Wilkie, A., Sandercock, P. A. G., & Wardlaw, J. M. (2006). Geographical access to imaging facilities for stroke patients in Scotland. *Health & Place*, 12(4), 617–630. doi:10.1016/j.healthplace.2005.05.004 PubMed

Shah, M. T., Bhattarai, S., Lamichhane, N., Joshi, A., LaBarre, P., Joshipura, M., & Mock, C. (2015). Assessment of the availability of technology for trauma care in Nepal. *Injury*, 46(9), 1712–1719. doi:10.1016/j.injury.2015.06.012 PubMed

Stephan, J. C. (1988). Technology assessment at the French National Centre for hospital equipment. *Health Policy (Amsterdam)*, 9(3), 341–347. doi:10.1016/0168-8510(88)90024-3 PubMed

Stratifikácia siete nemocníc. (2019). Ministerstvo zdravotníctva SR. <https://www.health.gov.sk/?izp-analyzy>

Suzuki, T., Isomi, Y., Tsuji, S., Tani, Y., Tanikawa, T., Yamasina, H., & Ogasawara, K. (2018). Penetration factors and introduction possibility for image diagnostic equipment. *Health Policy and Technology*, 7(2), 142–148. doi:10.1016/j.hlpt.2018.01.008

Tsekouras, K., Papathanassopoulos, F., Kounetas, K., & Pappous, G. (2010). Does the adoption of new technology boost productive efficiency in the public sector? The case of ICUs system. *International Journal of Production Economics*, 128(1), 427–433. doi:10.1016/j.ijpe.2010.07.041

Wei, Y., Yu, H., Geng, J., Wu, B., Guo, Z., He, L., & Chen, Y. (2018). Hospital efficiency and utilization of high-technology medical equipment: A panel data analysis. *Health Policy and Technology*, 7(1), 65–72. doi:10.1016/j.hlpt.2018.01.001

Winchester, M. S., & King, B. (2018). Decentralization, healthcare access, and inequality in Mpumalanga, South Africa. *Health & Place*, 51, 200–207. doi:10.1016/j.healthplace.2018.02.009 PubMed

Xing, Z., & Oyama, T. (2016). Measuring the impact of Japanese local public hospital reform on national medical expenditure via panel data regression. *Technological Forecasting and Social Change*, 113, 460–467. doi:10.1016/j.techfore.2016.07.026

Zuckerman, S., Hadley, J., & Iezzoni, L. (1994). Measuring hospital efficiency with frontier cost functions. *Journal of Health Economics*, 13(3), 255–280. doi:10.1016/0167-6296(94)90027-2 PubMed

KEY TERMS AND DEFINITIONS

Access to Health Care: In general, it is the distance to a particular health care provider. It contains four basic dimensions: physical availability, economic (affordability from the patient's perspective), non-discrimination (including time availability) and information availability.

Correspondence Analysis: Correspondence analysis is a dimension reduction technique which results are displayed on the map so called correspondence map. Plotted points on the correspondence map represent the relative numbers of the pivot table. The points' position reveals similarities between the different row categories, between the column categories, as well as the relationship between them. Correspondence analysis show hidden relationships and the association between the analysed variables.

Health System: Includes all activities that support, renew and maintain health. It represents the social system of all institutions, organizations, and resources whose main objective is to improve population health through targeted prevention, restoration and maintenance of population health.

Health Technology: Is an intervention used to promote health, prevention, diagnosis and treatment of acute or chronic diseases, or in rehabilitation, etc.

Magnetic Resonance Imaging Scans (MRI): Represents the most accurate spatial imaging of soft tissues of the investigated site that is a direct insight into the interior without surgical intervention with maximum accuracy. It allows very detailed views of organs and tissues throughout the body without the need for X-rays or ionizing radiation. MRI employs strong static and low frequency time-variable magnetic fields along with radio waves that, in a connection with computer processing, produce images showing injury, disease process, or the presence of abnormal tissues.

Process of Demographic Aging: A process influenced by the prolongation of human life expectancy and the transfer of strong population years (1970s and 1980s) to higher ages. By 2030, there will be a significant increase in the population of people over the age of 75, which means expectation of increased consumption of health and welfare due to the declining rate of self-sufficiency of the age.

Regional Disparities in Health (Health Inequalities): Represent differences, disparities, and inequalities in health between individuals and groups. They are determined by the individual social position (income, education, employment, natural predispositions - genetics, innate immunity, etc.).

Sustainability of the Health System: It represents primarily fiscal sustainability in the health care sector. The Slovak health care system is largely funded by wage payments, funding depends directly on labour market participation and growing earnings. It is also affected by the growth of the economic burden index due to the higher age of the population.

Section 16

Medical Technologies

Expanding Scope of Information Technology in Clinical Care

Gumpeny R. Sridhar

Endocrine and Diabetes Centre, Visakhapatnam, India

INTRODUCTION

Clinical medicine deals with data in many forms. Conceptually it is suited to gain from information technology (IT) in terms of gathering, analyzing and putting data to practical use. However there has been a chasm, with experts from the two fields often unable to find common ground. The system developed at the Endocrine and Diabetes Centre (EDC) had its origin nearly 30 years ago; it was an end-user initiative followed by features addition by an IT expert. This is different from most earlier attempts where experts from the two fields tried to understand the needs and possibilities of each other with attendant bottlenecks in implementation.

When the EDC was set up 30 years ago, documentation was among the first priorities. Without losing focus of the primary aim, viz, patient care, the kind and quality of data was slotted to be developed into EMR format. Where the knowledge of the clinician was not sufficient to go ahead, collaboration with a software professional was sought; the latter could take over without glitches, because the product, developed until then, was geared in logic and technology to be taken over to the next stage. Importantly the evolution of the IT infrastructure was achieved without any interruption in patient care.

To our knowledge this is the first doctor-initiated EMR system, which was developed and evolved over nearly 30 years. The purpose of this chapter is to document its genesis and evolution and to point to future areas where it can be used. In its current iteration the project is being utilized at the point of care without disrupting the work flow, while retaining the advantages of data digitization and for allowing for addition of further features.

Clinical medicine aims to provide comprehensive care, which is personalized and predictive. As synergy between information technology and clinical encounters led to the electronic medical record (EMR) being the fulcrum of the health care system. What began as a 'electronic paper records' (Price M, Singer A & Kim J, 2013), has expanded to a much broader spectrum to be gathered, collated and utilized, adding digital images, sharing in real time, with the possibility in the near future to integrate other data including genomics and proteomics. In this chapter the evolution of EMR system at the Endocrine and Diabetes Centre is presented along with a blueprint of how it is planned to expand integrating future technologies and processes.

BACKGROUND

Conventionally, clinical medicine deals with data, obtained as non-linear, non-digital form. Paper medical records formed the bedrock for much of initial clinical research and progress. Despite their advantages re-use of data is not practical (Sridhar GR, Rao AA, 2009).

DOI: 10.4018/978-1-7998-3479-3.ch131

However in the management of non-communicable diseases such as diabetes mellitus, obesity, hypertension and coronary artery disease, a collaboration between the health care team and the patient is essential, with the patient sharing responsibility in self-care, outside of the physician's office. The rapid applicability of IT allows a practical means to do so. In addition newer digital technologies result in generation of large amount of data, which must be analyzed and used appropriately (Cahn A Akirov A& Raz I. 2017). EMR system has the many advantages of storing nearly unlimited data, with the ability to access, analyze and present it in a format that is easy to interpret, along with providing accurate and legible notes which can track the individual's lifetime of health status, medicine use, laboratory results, images from many sources, transferable across locations. They also provide clinical guidelines, flag abnormal results, remind tests to be done, and reduce medication errors (Sridhar GR, Rao AA et al 2009). In addition newer technologies allow remote consultation with physicians, electronic prescription and drug delivery along with feedback about medicine use (Poon EG et al 2006).

The initial hesitation of using EMR in the clinic, where, in the early stages, the physician seems to pay more attention to the computer or tablet screen than to the person sitting in front of her, is overcome with the many advantages of digital data (Doyle RJ, Wang N et al 2012). Integrating with older paper version of medical records is another major barrier for conversion into EMR, along with software issues in coding, customization, hardware maintenance, security and confidentiality (Pradeepa R, Prabu AV et al 2011). Newer skills are required to not lose emotional and eye contact with the patient. Essentially the EMR system must take into account the workflow so that it aids rather than hamper patient care. The bottom line is often the reluctance of physicians to use the newer systems. Despite all these, barriers not only from the end-users, but other stakeholders also, Serrano et al have concluded recently that health information systems could be beneficial in improving patients' quality of life and of care (Serrano A, Guzman JG et al 2018).

The EMR is ultimately integrated into a health information technology which includes decision-support systems which leads to cost saving at the system level (Ali MK, Shah S& Tandon N 2011).

The aim is to employ digital tools for individualized and upgraded delivery of medical care. Digital medicine has the potential to improve patient satisfaction, lower the cost of care, and enable clinical research not confined to merely academic centres (Stenholm SR & Topol EJ 2017). While information technology is ubiquitous, medical field has been untouched, except in isolated centres (Sridhar GR, Murali G 2011, Pradeepa R, Prabhu V et al 2011). Riddle et al (Riddle MC, Bakris G et al 2018) summarized the limitations of IT in the clinical practice: a general lack of harmony in the way data is collected and managed.

FOCUS OF THE ARTICLE

Evolution of the EMR system at Endocrine and Diabetes Centre (EDC)

When the EDC was established in 1988, patient records were maintained as text using paper files. With the aim of future digitized entry, the data was categorized into Identification (unique number for each new subject), Name, family name, gender, age, address, city and zip code. A second section related to the history of the patient, followed by the following sections on physical examination, investigations, treatment and follow up.

After two more years, the kind of (endocrine) diseases with which the patients presented was reviewed as well as the symptoms and signs for each. Data was separated into categories (eg in diabetes: duration

of diabetes, onset age, appetite, change in weight, weakness, shortness of birth etc); the main groups of diseases could be categorized into diabetes, thyroid, short stature, obesity, infertility and others (other disorders not fitting into the above categories).

Beginning of EMR

Next a basic EMR system was developed using database management system, having separate files for each condition (listed above). Each field in the feature-wise entry of the paper system was converted to a separate field in the database.

Down the line, Mr Venkat Yarabati who provided software support opined that although the next platform, Paradox was good, migration to a more recognized system was preferable, which is less likely to be taken off the market. Eventually his hunch proved correct.

The Next Step

To develop integrated software in MS Access, the disparate files (diabetes, thyroid etc) were all put together with three different files displayed together on the screen. Auto-generation of ID numbers as well as auto-save of data.

One file, common to all patients had the ID number, name, gender, age and address. A drop down menu provided a list of primary conditions, the second set of files, for which the patient presented (viz diabetes, thyroid, short stature, obesity, infertility and others). The third file at the bottom of the screen had fields for follow up data.

Calculated fields were added (viz, onset of diabetes by subtracting duration of diabetes from current age, body mass index from the height and weight fields etc). Besides, certain fields were blanked out depending on the gender or suspected condition (eg field on menstrual cycle was blanked for men; symptom and sign fields were appropriately presented or blanked for conditions of thyroid excess or deficiency).

In the follow up, only relevant portions of the first visit entries were displayed, along with data for a horizontal one-line entry at each visit. Feedback from the clinician is essential in arriving at these decisions. There was a periodical review and re-calibration of the system. To allow entry of other information, memo fields were included. Chronological display of important data provides a clearer picture of the course of disease and helps the clinician in taking corrective actions.

A printout of the generated form was given to the (new) patient at the end of the consultation. Depending on the data inputs, automatic advice was provided based on automated system generated print out, based on the rules incorporated appropriately (eg to stop smoking if the patient was a current smoker; to take care of the feet if there were symptoms or signs of peripheral neuropathy, or if the duration of diabetes was over 10 years, which is a risk factor, or if there were abnormal signs of the feet). Fields were also filled for the diagnosis (type 1 or type 2 diabetes, hypertension if the patient was on antihypertensive medicines, ischemic heart disease if the patient provided a history, different varieties of dyslipidemia depending on the values of the lipids in the fields for lab investigations. Similarly a numeric range was provided with a flag if the number exceeded it to avoid data entry mistakes.

An important first decision was taken to enter data in numeric fields even for qualitative data, because that would allow for analysis later on without transformation. Since only one person used the system initially, information was entered as categorical data: a number was assigned for differences in each symptom (eg '1' for normal appetite, '2' for increased appetite, '3' for decreased appetite etc for different variables).

The Next Step in the Evolution of the Database

Until this stage, all the data was entered by the clinician. In addition the prescriptions were hand-written. To improve the work-flow, a local area network was devised so that basic patient met the physician.

The Current Phase

The following software were used in the development of the current system: SQL Server 2012 as backend database, MS Access as front end, Access Reports for reporting and Microsoft Office Chart.

Password protected access to the EMR was devised. Those at the reception were provided access to adding a new patient along with her details, to enter the basic anthropometric and laboratory parameters, and the option to search whether the patient had registered earlier (based on the name and or family name) to avoid duplication. Mr Sairam worked on this aspect of the project.

This improved the efficiency of patient turn-around where routine tasks and entry of data are either automated or out-sourced at the reception so that the time of the patient and the physician are more efficiently utilized.

Practical Caveats

There were some practical lessons that were learned. Initially, the ID numbers were written manually and then entered into the EMR. This was error-prone and duplicate records were generated. From the outset an iron-clad procedure was followed for back up of data (every day initially, every week, later) on a hard-disk that is stored in a different physical location. It is very important to consider the possibility of hacking attacks or ransom-ware and therefore the system is best isolated from general internet.

Currently the data is stored in a separate server not connected to the internet, and all data being mirrored on a second hard disk in the server.

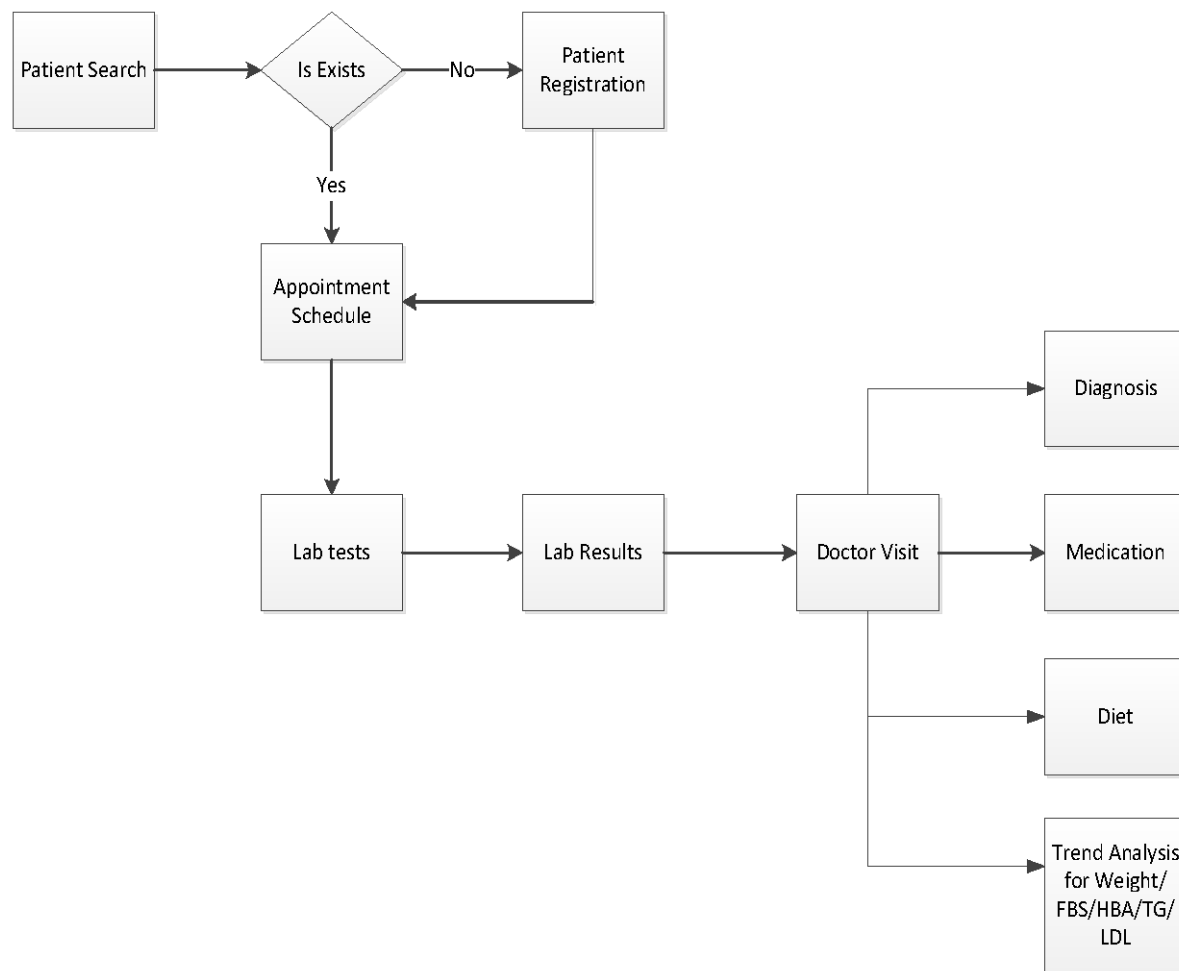
It is essential to bring across all users of the system together, and interact with the software developers. The difficult part is not for the software to be structured; difficulty arises in the health-care team providing the data that must be gathered and in which sequence. This needs effort, because traditional note taking entails non-structured free hand entry. Sufficient time must be provided for a balance to be struck between them. Without this, trying to put the system in place is likely to be fruitless.

On a literature search, the current system appears to be based on the database developed at the St Thomas Hospital, London (Carson ER, Carey S et al 1990). During the early 1970's when the London database was first begun, data was acquired by structured questions and later batch-entered into a main-frame computer. Twenty years later, the personal computer made it possible to both capture and record data simultaneously. Physical limitations of the hardware thirty years ago resulted in the presentations of data from 4,500 patients in comparison to the current system at EDC handling with data from more than 80,000 patients, and growing by the day.

Algorithms used for calculation of following items: body mass index, auto population of DX for Type 2 diabetes, categorization of diagnostic lipid disturbances diet prescription, calculation, patient search based on EDC Id, name, gender, city, mobile etc, patient appointment and fee calculator.

The workflow is provided in Figure 1

Figure 1. Workflow for the EMR at EDC.



SOLUTIONS AND RECOMMENDATIONS

With the system in its current iteration comprehensive and robust for clinical use, it stands ready to be interactive with and inclusive of data from other sources. An integration with the robotic biochemical analyzer allows for bar-coded samples to be analyzed and the results displayed on the physician's desk in real-time. With this and other devices for potential integration, collaboration between the instrument supplier and the software specialist is essential to allow transfer of data. Similarly an integration with the pharmacy the prescription can be sent ahead and the medicines delivered to the patient either by hand, or may be sent by post. A reconciliation software in the pharmacy will keep track of the stock available and send in automatic orders for medicines which are close to a pre-determined critical value.

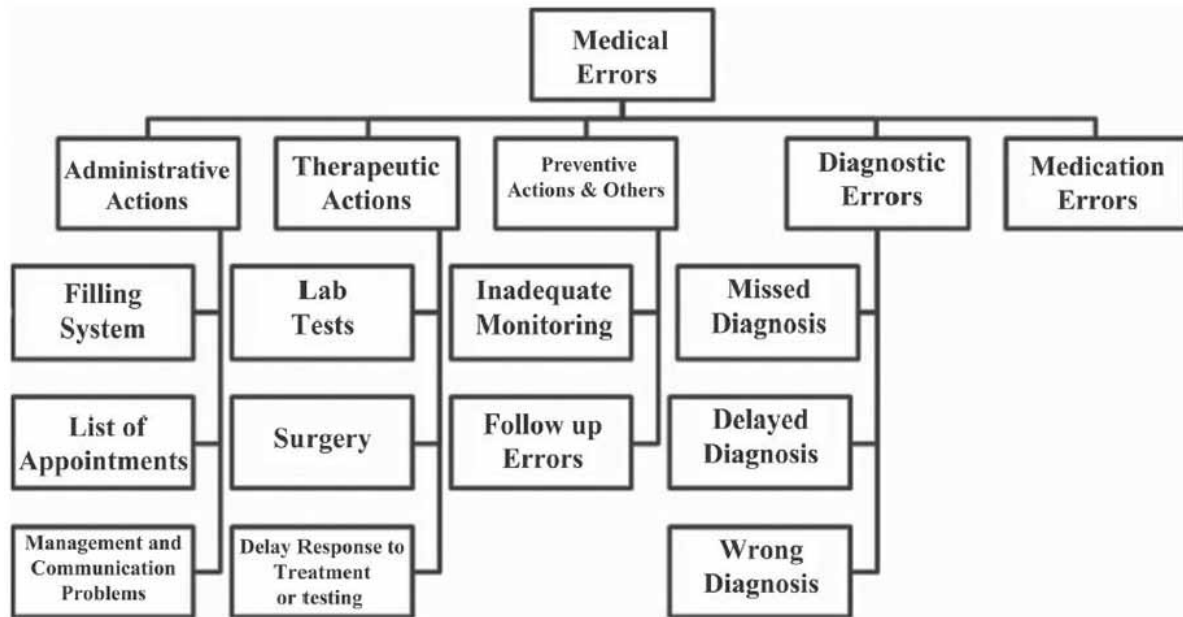
Other digital data can be included in the patients EMR such as images of the eye fundus employing a digital fundus camera, thermal images of the feet, electrocardiographic tracings, ultrasound images, x-rays and clinical photographs. Images and recordings from all these sources can be captured, transmitted and stored in digital format.

Information from different sources can be integrated and retrieved in the patient's EMR . This saves time and effort of the physician while providing value for the cost of capturing data.

On follow up it will be possible to have a graphical representation of different parameters – individually or collectively.

Essentially, EMR allows for the capture and utilization of information and prevent medication errors to occur (Figure 2): patient identification errors, prescription error such as wrong drug, dosage and incorrect dose calculation, failure to adjust dosages in patients based on the patient allergies

Figure 2. Errors that can be prevented by EMR.



QUALITY OF CARE

The goal of these innovations is to improve patient care and outcomes. Quality can be measured on different domains such as clinical parameters, biochemical parameters along with economic and psychologic aspects (Kapur A 2001). It is important to differentiate between documentation and outcome, viz performing an action (biochemical test(s), clinical evaluation) versus improved outcome as a result of the action (Rao AA & Sridhar GR 2007. Iljaz R, Brondik A, Zrimec T & Cukjati I, 2017). A recent review suggested that decision-support tools having detailed features, interaction with patients accessible to them give better outcomes (Ali MK, Shah S& Tandon N 2011).

With the aim of improving coordination among health care team members by reducing missing clinical information, and providing support in clinical procedures, there is little real world evidence that EMRs truly improve outcomes in primary care settings. Crosson et al have shown that the mere employment of EMRs in primary care practices is not enough for improving quality of care (Crosson JC, Strickland PAO et al 2007). A synergy between EMR vendors and physicians using the EMR systems should be an ongoing process.

A literature review published in 2008 by Adaji et al, based on 24 articles that satisfied the pre-set criteria (out of a pool of 444 articles) showed that the use of information technology in diabetes care has the potential to improve diabetes care by improving interaction between the patient and the care

providers (Adaji A et al 2008). Even here, improvement in process measures was more than in outcome measures. Some of the impeding factors were, having to learn new skills, lack of adequate time, a fear of using computers and anxiety results from the need for change. These were overcome by training and by integrating the system into the usual process of care (Adaji A et al 2008).

On a more limited scale, healthcare information technology improved adherence to taking medicines from a review of 13 randomized clinical trials where electronic interventions were performed (Misono AS et al 2010). A critical aspect to success required a ‘substantial degree of patient motivation at the onset’ (Misono AS et al 2010). However the quality of studies were variable in rigour and duration.

Ciemins et al studied 495 adults with diabetes, initially provided with education to the primary care provider and. This was followed by EMR implementation (Ciemins et al 2009). Factors for success lay at the level of the system and at the provider. The intervention improved both processes and outcomes: at the end of the study, subjects were more likely to have received preventive screening, and to have better control of clinical risk factors. The system was developed and implemented ‘*to fit into tailored office workflow process*’ (Ciemins et al 2009).

A large quality related advantages of EHRs over traditional paper records was published from a large number of adults with diabetes (n:27,707) from 46 office practices (Cebul RD et al 2011). Following adjustment for covariates, there was substantially higher achievement of composite standards for care of diabetes among sites using EHR by 35.1 percentage points. Similarly, the composite standards achieved for outcomes was 15.2 percentage points higher in the EHR using sites. These were associated with improvement in care and outcomes as well (Cebul RD et al 2011). The authors concluded that appropriate use of EHRs may result in improvement in the quality of care in the population studied. Part of the reason for their better outcomes with EHR usage could be due to better quality of clinical decision support. The decision support system was not generic, but was targeted to improve standards of care.

A recent study from Slovenia added to the evidence that innovative approach using e-healthcare resulted in better self-management in subjects with diabetes who are not utilizing insulin for glycemic control (Iljaz R, Brodnik A, Zrimec T, Cukjati I 2017).

COST BENEFIT ANALYSIS

An AHRQ Publication on the costs and benefits of health information technology (HIT) prepared for the Agency for Healthcare Research and Quality (2006) analyzed 256 published studies on the benefits and costs of health information technologies. The cost-benefit analysis predicted that implementation of HIT could result in significant savings of cost. The benefits were projected to overweight the cost of investment.

Since implementing an electronic health record system has costs, not necessarily financial, the benefits could also relate to issues such as productivity, efficiency, decision-making and compliance to guidelines.

Barriers were related to situational, cognitive, physical, knowledge and attitudinal.

A more recent review on economic evaluation of health information systems was carried out based on publications between January 2000 and June 2012 (n:33). The authors concluded that concluded that economic evaluations were made with differences in methods, metrics and assumptions. However there was ‘some evidence of value for money in selected healthcare organizations and HIS types’ (Bassi J & Lau F. 2013). The caveat made by the authors that there was need for high quality studies of economic evaluation, with regard to six key components: ‘having a perspective, options for comparison, time frame, costs, outcomes, and comparison of costs and outcomes for each option’ (Bassi J & Lau F. 2013).

FUTURE RESEARCH DIRECTIONS

Merging and Integration of Data

While traditional data from clinical encounters and laboratory and financial is currently available, one must plan for the increasing availability of data from other sources such as genomics, proteomics and metabolomics. Improved technologies and plunging prices presage large amounts of data from newer technologies to be merged with clinical care. Issues of analysis, interpretation, usage and confidentiality are bound to arise.

Precision Medicine

Precision medicine refers to “..an approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle of each person” [<https://ghr.nlm.nih.gov/primer/precisionmedicine/definition>]. It requires integrating genomics and electronic medical records via research in basic sciences, clinical sciences and translational science (Ritchie MD et al 2015). Bertsimas et al reported that when personalised, diabetes management improved metabolic outcomes compared to the standard of care (Bertsimas D, Kallus N, Weinstein AM, Zhuo YD 2017). Thus the EMR will form the fulcrum where new data streams feed for a more comprehensive and individualized treatment. In the current phase, integration will be required across electronic record systems and genetic datasets, which differ in different medical centres. Reconciling these diversities will be a major problem awaiting solution.

Once the goal is met, physicians at their point of can access relevant genetic information in the EMR, which allows better diagnosis and resultant targeted care to patients (Ritchie MD et al 2015).

Issues of Consent and Privacy

Privacy and consent must be addressed in the march toward precision medicine (Sridhar GR, 2018). Often, consent is taken for a limited purpose in studies involving genomics; with further tests genomic data can be linked with EMR. The issue of consent if taken on paper becomes contentious. Therefore online consent obtaining procedures were suggested (Kogetsu A et al 2018).

Mutual trust between the investigator and the subject is essential. Considering the limitations of paper-based obtaining of consent, new methods were proposed where dynamic consent can be given online, ensuring interaction and continuity (Budin-Ljosne et al 2017). This allows subjects to either alter or remove their consent.

Issues of security in accessing data is a critical component in eHealth or the ‘use of information and communication technology for health’ (Budin-Ljosne et al 2017). Subject data can be encrypted, but authentication for retrieval needs equal attention. Currently, passwords are employed for authentication, which is prone to data leaks. Authentication of genuine user is required at the time of registration and at each log in.

For authentication, three broad methods may be used: ‘What one knows’, ‘What one has’ and ‘What one is’ ((Budin-Ljosne et al 2017). Authentication by what one knows may be a password or secret question, but the security is not sufficient. What one has may be a device or a token to generate a password, which is both inconvenient and prone to loss or theft. The third or what one is could be authentication, using biometry such as fingerprints or voice. It is not preferred as a single authentication method.

Currently, a two-step authentication method is preferable, consisting of a first-step password or equivalent, followed by the second step device or one time password (Budin-Ljosne et al 2017).

Defining Conditions and Diseases

Data is not often so clean and narrow in numeric form as in the EDC database; genomic data is linked to patient information stored in EMRs spread over institutions containing various types of data. To link genetic information with disease requires methods to define or at least describe the condition found in the EMR (Robinson JR et al 2018). Such efforts are worthwhile because longitudinal data of patients progress is available in EMRs, which can be linked to genomic and related data. Merging the two provides a fine tuned and nuanced picture of the phenotype-genotype correlation.

Before EMRs, phenotypes were curated from structured and unstructured data from a variety of sources including computerized billing codes, lab results, vital signs, use of drugs. Diagnostic coding of diagnosis provides a coarse-grained information, but suffers from changing coding criteria and errors in assigning codes.

The richest source of information is available in clinical documentation. Therefore, narrative text without uniformity must be processed using tools such as the natural language processing (NLP). General-purpose NLP systems are usable across a number of conditions, but they suffer from lower recall and precision compared to purpose-built approaches. For text-based annotation, NLP methods can identify sentence structure (Budin-Ljosne et al 2017).

Many methods were utilized in curating phenotypes: rule based, natural language processing, machine learning (including learning with noisy data, unsupervised phenotype discovery, hybrid approaches) and collaborative frameworks. A review of these methods was published recently (Banda JM et al 2018). *Rule-based* methods are appropriate where a clear diagnosis or procedure codes is available. *Phenomics* refers to the transdisciplinary field where genomic information is linked to phenotype data extracted from EMRs. It has a limited applicability when dealing with complex phenotypes or with datasets that are less standardized.

Where data are recorded in unstructured form as in discharge notes, radiology reports, pathology reports even in EMRs, information can be extracted by employing pattern-matching techniques against standard vocabularies (Hersch WR & Greenes RA 1990).

Currently, NLP is the common tool to extract electronic phenotyping; it is used along or along with rule-based techniques or statistical learning methods. *Symbolic techniques* depend on predefined semantic relationships, but with limited versatility. *Statistical model* is currently used more extensively in text processing. Both conceptual and contextual cues are utilized for phenotyping (Banda JM et al 2018). NLP is expected to play a prominent role in phenotyping efforts to both recognizing phenotypes as well as to process text and extract features needed for statistical learning.

Huang et al published a machine learning method to phenotype subjects with diabetes (Huang et al 2007). They employed feature selection via supervised model construction (FSSMC). The technique ranked 47 manually distilled features from an initial set of 410 structured variables. Using machine learning classifiers, naïve Bayes, C4.5 and IBI, those with highest performance were selected.

Support vector machines (SVMs) were employed for rheumatoid arthritis, an autoimmune disease. Associated rule mining method generated rules understandable by humans employing demographics, lab results and coded billing data (Banda JM et al 2018).

Other Sources of Data and Predictive Analysis

In addition to data from genomics, proteomics and metabolomics, data from other sources can also be integrated into the EHR: viz, vital signs (pulse rate, blood pressure, continuous glucose monitors, activity and sleep). A rich collection of phenotypic and genomic data would allow for machine learning techniques to predict the outcomes (Ritchie MD, 2018a). Practical demonstration of such use is available for predicting early eye changes in diabetes (Imani et al 2015). Proof of concept work was published in utilizing clinical variables to predict psychological outcomes (Narasinga Rao MR Sridhar GR ...et al 2010).

CONCLUSION

Clinical data lay scattered and often un-analyzable. Technology allowed the data to be organized in a manner that allowed it to be put to a lot more use than the original intention. With time the amount of information has burgeoned, requiring for a new eco-system to deal with it using information technology, workflow and methods to synthesize and utilize the data. Current improvements from clinical processes to outcomes must be addressed (Ali MK, Shah S& Tandon N 2011). Predictive modeling of diabetic complications and risk profiling is possible from analyzing data from electronic health systems (Liu B, LI Y et al, 2018; Marini S, Dagliati A et al 2016). Advances in combining clinical and genomic data (Kaiser J, 2018) helps in the transfer from the bench to the bedside (Sridhar GR & Murali G 2011). Integrating innovative advances in technologies into the health care system appears a key component in dealing with increasing health care costs and in an approach toward the goal of universal health care (Blood DE et al 2018).

REFERENCES

- Adaji, A., Schattner, P., & Jones, K. J. (2008). The use of information technology to enhance diabetes management in primary care: A literature review. *Informatics in Primary Care*, 16, 229–237. PMID:19094410
- Ali, M. K., Shah, S., & Tandon, N. (2011). Review of electronic decision-support tools for diabetes care: A viable option for low- and middle-income countries? *Journal of Diabetes Science and Technology*, 5(3), 553–570. doi:10.1177/193229681100500310 PMID:21722571
- Banda, J. M., Semeviratne, M., Boussard, T. H., & Shah, N. H. (2018). Advances in electronic-phenotyping: From rule-based definitions to machine learning models. *Annual Review of Biomedical Data Science*, 1(1), 12–56. doi:10.1146/annurev-biodatasci-080917-013315 PMID:31218278
- Bassi, J., & Lau, F. (2013). Measuring value for money: A scoping review on economic evaluation of health information systems. *Journal of American Informatics Association*, 20(4), 792–801. doi:10.1136/amiajnl-2012-001422 PMID:23416247
- Bertsimas, D., Kallus, N., Weinstein, A. M., & Zhuo, Y. D. (2017). Personalised diabetes management using electronic medical records. *Diabetes Care*, 40(2), 210–217. doi:10.2337/dc16-0826 PMID:27920019
- Bloom, D. E., Khoury, A., & Subbaraman, R. (2018). The promise and peril of universal health care. *Science*, 361. PMID:30139847

- Budin-Ljøsne, I., Teare, H. J. A., Kaye, J., Beck, S., Bentzen, H. B., Caenazzo, L., Collett, C., D'Abramo, F., Felzmann, H., Finlay, T., Javaid, M. K., Jones, E., Katić, V., Simpson, A., & Mascalcioni, D. (2017). Dynamic Consent: A potential solution to some of the challenges of modern biomedical research. *BMC Medical Ethics*, 18(1), 4. doi:10.1186/12910-016-0162-9 PMID:28122615
- Cahn, A., Akirov, A., & Raz, I. (2018). Digital health technology and diabetes management. *Journal of Diabetes*, 10(1), 10–17. doi:10.1111/1753-0407.12606 PMID:28872765
- Carson, E. R., Carey, S., Harvey, F. E., Sonksen, P. H., Till, S., & Williams, C. D. (1990). Information technology and computer-based decision support in diabetic management. *Computer Methods and Programs in Biomedicine*, 32(3-4), 179–188. doi:10.1016/0169-2607(90)90100-N PMID:2249417
- Cebul, R. D., Love, T. E., Jain, A. K., & Hebert, C. J. (2011). Electronic health records and quality of diabetes care. *The New England Journal of Medicine*, 365(9), 1–9. doi:10.1056/NEJMs1102519 PMID:21879900
- Ciemins, E. L., Coon, P. J., Fowles, J. B., & Min, S. J. (2009). Beyond health information technology: Critical factors necessary for effective diabetes disease management. *Journal of Diabetes Science and Technology*, 3(3), 452–460. doi:10.1177/193229680900300308 PMID:20144282
- Crosson, J. C., Strickland, P. A. O., Hahn, K. A., Bloom, B. D., Shaw, E., Orzano, A. J., & Crabtree, B. F. (2007). Electronic medical records and diabetes quality of care: Results from a sample of family medical practices. *Annals of Family Medicine*, 5(3), 209–215. doi:10.1370/afm.696 PMID:17548848
- Doyle, R. J., Wang, N., Anthony, D., Borkan, J., Shield, R. R., & Goldman, R. E. (2012). Computers in the examination room and the electronic health record: Physicians' perceived impact on clinical encounters before and after full installation and implementation. *Family Practice*, 29(5), 601–608. doi:10.1093/fampra/cms015 PMID:22379185
- Hersch, W. R., & Greenes, R. A. (1990). SAPHIRE: An information retrieval system featuring concept-matching, automatic indexing and probabilistic retrieval. *Computers and Biomedical Research, an International Journal*, 23, 405–420. PMID:2225787
- Huang, Y., McCulagh, P., Black, N., & Harper, R. (2007). Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine*, 41(3), 251–262. doi:10.1016/j.artmed.2007.07.002 PMID:17707617
- Iljaz, R., Brondik, A., Zrimec, T., & Cukjati, I. (2017). E-healthcare for diabetes mellitus type 2 patients – a randomized controlled trial in Slovenia. *Zdr Varsity*, 56, 150–157. PMID:28713443
- Imani, E., Javidi, M., & Pourreza, H. R. (2015). Improvement of retinal blood vessel detection using morphological component analysis. *Computer Methods and Programs in Biomedicine*, 118(3), 263–279. doi:10.1016/j.cmpb.2015.01.004 PMID:25697986
- Kaiser, J. A. (2018). A new portal for patient data. *Science*, 361(6399), 212. doi:10.1126/science.361.6399.212 PMID:30026205
- Kapur, A. (2001). Influence of socio-economic factors on diabetes care. *International Journal of Diabetes in Developing Countries*, 21, 77–85.

Kogetsu, A., Ogishima, S., & Kato, K. (2018). Authentication of patients and participants in health information exchange and consent for medical research: a key step for privacy protection, respect for autonomy, and trustworthiness. *Frontiers in Genetics*, 9, 167. doi:10.3389/fgene.2018.00167

Liu, B., Li, Y., Ghos, S., Sun, Z., Ng, K., & Hu, J. (2018). *Simultaneous modeling of multiple complications for risk profiling in diabetes care*. arXiv:1802.06476v1 [cs.Lg]

Marini, S., Dagliati, A., Sacchi, L., & Bellazzi, R. (2016). Learning T2D evolving complexity from EMR and administrative data by means of continuous time Bayesian networks. *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016)*, 5, 338-344. 10.5220/0005708103380344

Misono, A. S., Cutrona, S. L., Choudhry, N. K., Fischer, M. A., Stedman, M. R., Liberman, J. N., ... Shrank, W. H. (2010). Healthcare information technology interventions to improve cardiovascular and diabetes medication adherence. *The American Journal of Managed Care*, 16, SP82–SP92. PMID:21314228

Narasinga Rao, M. R., Sridhar, G. R., Madhu, K., & Appa Rao, A. (2010). A clinical decision support system using multi-layer perceptron neural network to predict quality of life in diabetes. *Diabetes & Metabolic Syndrome*, 4(1), 57–59. doi:10.1016/j.dsx.2009.04.002

Poon, E. G., Blumenfeld, B., Hamann, C., Turchin, A., Graydon-Baker, E., McCarthy, P. C., Poikonen, J., Mar, P., Schnipper, J. L., Hallisey, R. K., Smith, S., McCormack, C., Paterno, M., Coley, C. M., Karson, A., Chueh, H. C., Van Putten, C., Millar, S. G., Clapp, M., ... Broverman, C. A. (2006). Design and Implementation of an Application and Associated Services to Support Interdisciplinary Medication Reconciliation Efforts at an Integrated Healthcare Delivery Network. *Journal of the American Medical Informatics Association*, 13(6), 581–592. doi:10.1197/jamia.M2142 PMID:17114640

Pradeepa, R., Prabu, A.V., Jebarani, S., Subhashini, S., & Mohan, V. (2011). Use of a large diabetes electronic medical record system in India: clinical and research applications. *Journal of Diabetes Science and Technology*, 5, 543-52.

Price, M. (2013). Adopting electronic medical records. *Canadian Family Physician Medecin de Famille Canadien*, 59, e322–e329. PMID:23851560

Rao, A. A., & Sridhar, G. R. (2007). Quality of care: Assessment. *Lipids in Health and Disease*, 6(1), 12. doi:10.1186/1476-511X-6-12 PMID:17442124

Riddle, M. C., Bakris, G., Boulton, A. J. M., Blonde, L., D'Alessio, D., Greene, E. L., de Groot, M., Hood, K., Hu, F. B., Kahn, S. E., Kaul, S., LeRoith, D., Moses, R. G., Rich, S. S., Rosenstock, J., Tamborlane, W. V., Wylie-Rosett, J., & Reynolds, L. (2018). Big topics for Diabetes Care in 2018: Clinical guidelines, costs of diabetes and information technology. *Diabetes Care*, 41(7), 1327–1329. doi:10.2337/dci18-0035 PMID:29934474

Ritchie, M.D. (2018a). Large-scale analysis of genetic and clinical patient data. *Annual Review of Biomedical Data Science*, 1, 263-274.

Ritchie, M. D., Andrade, M. D., & Kulvaniemi, H. (2015). The foundation of precision medicine: Integration of electronic health records with genomics through basic clinical, and translational research. *Frontiers in Genetics*, 6, 104. doi:10.3389/fgene.2015.00104 PMID:25852745

Robinson, J. R., Wei, W. Q., Roden, D. M., & Denny, J. C. (2018). Defining phenotypes from clinical data to drive genomic research. *Annual Review of Biomedical Data Science*, 1(1), 69–92. doi:10.1146/annurev-biodatasci-080917-013335

Serrano, A., Guzman, J. G., Xydopoulos, G., & Tarhini, A. (2018). *Analysis of barriers to the deployment of health information systems: a stakeholder perspective*. Information Systems Frontier. doi:10.1007/10796-018-9869-0

Sridhar, G. R. (2018). Encode, decode and diabetes. In Cognitive science and health bioinformatics. Advances and applications. Springer. doi:10.1007/978-981-10-6653-5_4

Sridhar, G. R., & Murali, G. (2011). Computerization of data in diabetes centers. *International Journal of Diabetes in Developing Countries*, 31(2), 48–50. doi:10.1007/13410-010-0004-0

Sridhar, G. R., Rao, A. A., Muraleedharan, M. V., Kumar, R. V. J., & Yarabati, V. (2008). Electronic medical records and hospital management systems for management of diabetes. *Diabetes & Metabolic Syndrome*, 3(1), 55–59. doi:10.1016/j.dsx.2008.10.008

Stenholm, S. R., & Topol, E. J. (2017). Digital medicine, on its way to being just plain medicine. *Npj Digital Medicine*, 1, 5.

ADDITIONAL READING

MIT Critical data. (2016). *Secondary analysis of electronic health records*. Springer.

Rinaldi, G. (Ed.). (2017). *New perspectives in medical records*. Springer. doi:10.1007/978-3-319-28661-7

Shortliffe, E. M., & Cimino, J. J. (2006) Biomedical informatics. Computer applications in health care and biomedicine. Springer NY doi:10.1007/0-387-36278-9

KEY TERMS AND DEFINITIONS

Diabetes: A disease in which the body does not have enough of the hormone insulin.

Endocrine: Related to chemical secretions called hormones that serve as messengers to neighboring or far off cells.

Genomics: Relating to the study of genes, which essentially direct the function of cells in the body's cells.

Proteomics: Study of structure of proteins, which are produced as a result of the directions given by the genes.

Signs: The findings observed or elicited by the physician in a sick person.

Symptom: The complaints made by the patient as a result of her illness.

Artificial Neural Networks in Medicine: Recent Advances

16

Steven Walczak <https://orcid.org/0000-0002-0449-6272>*University of South Florida, USA*

INTRODUCTION

Medicine is a field closely coupled with and producing big data (Najafabadi et al., 2015), especially with the growing adoption of electronic health records (EHRs) in the United States (Bourgeois & Yaylaci, 2010; Mennemeyer et al., 2016) and the world (Wager et al., 2013, Appendix C). Medical big data serves as a critical resource for medical research and clinical decision making. Artificial neural networks (ANNs), along with other machine learning approaches, have been shown to be an effective method for analyzing medical big data to develop diagnostic and prognostic systems (Pastur-Romay et al., 2016).

Although ANNs have a short history of application in the field of medicine, with the first published research appearing in 1990 (Asada et al., 1990; Baxt, 1990; Dassen et al., 1990), there has been a continuing and growing trend of research investigating ANNs in medicine. Interestingly, two of the three articles published in 1990 were published in medical journals, while only Baxt's (1990) article on diagnosing heart attacks (myocardial infarctions) was published in an information technology journal, but he followed this work up with multiple publications in medical journals (Baxt, 1991, 1992). Searching the National Library of Medicine PubMed database for articles containing the term "artificial neural network" combined with any of the terms: medicine, medical, hospital, patient, diagnosis, prognosis, clinic, or pharma ([https://www.ncbi.nlm.nih.gov/pubmed/?term="artificial+neural+network"+AND+\(medicine+OR+medical+OR+hospital+OR+patient+OR+diagnosis+OR+prognosis+OR+clinic+OR+pharma\)](https://www.ncbi.nlm.nih.gov/pubmed/?term="artificial+neural+network"+AND+(medicine+OR+medical+OR+hospital+OR+patient+OR+diagnosis+OR+prognosis+OR+clinic+OR+pharma))) produces the results shown in Figure 1, which displays the annual increase in articles focusing on ANNs in medicine.

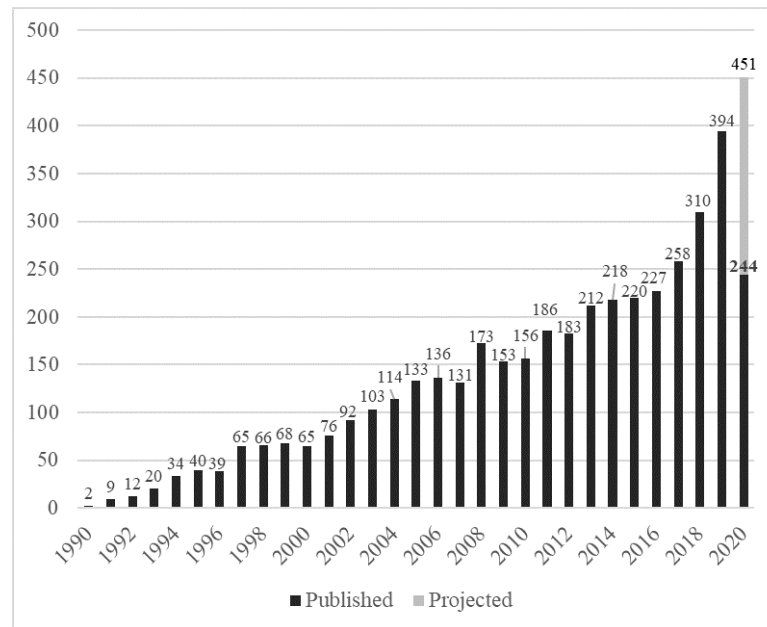
The amount of ANN in medicine research is even greater than depicted in Figure 1, since terms like "deep learning neural network" or "convolution network" or "evolutionary network," which are all types of ANN, are not included. Other researchers have also indicated a larger quantity of ANN research in medicine, with 473 articles in 1998 (Dybowski, 2000) as opposed to the 66 articles identified with the more specific search criteria. Other research claims an earlier start date, 1981, for the beginning of ANN usage in medical decision support (Miller, 1994), but here we report only those articles listed in the PubMed database.

ANN research in medicine is used to develop models and systems for a variety of applications including: decision support systems for both patients and surgeons, diagnosis, prognosis, resource planning and allocation, and variable significance and protocol heuristic evaluation. The goal of any research in the field of medicine should primarily be to improve the quality of life of the patient and secondarily to promote workflow efficiencies and cost reductions.

Next a brief definition of ANNs and a short historical perspective is presented. The purpose of this article is not to instruct researchers in how to develop an ANN, but rather to examine recent trends and

DOI: 10.4018/978-1-7998-3479-3.ch132

Figure 1. Medline/PubMed ANN articles in medicine as of July 15, 2020



advances in ANN research in medicine. Readers interested in a more general introduction to ANNs should examine the readings listed in the Additional Readings section. After the brief definition and historical background, recent research is described and trends in applying ANNs in medicine are discussed. Finally, future research directions are presented.

BACKGROUND

Artificial neural networks (ANNs) are a machine learning based classification and forecasting tool, based on modeling the neuronal activity of the human brain. ANNs come in a variety of architectures, which are represented by a large collection of interconnected processing elements called neurodes, and utilize a wide range of machine learning algorithms. These algorithms may be classified as either unsupervised learning or supervised learning or hybrid learning algorithms that utilize both unsupervised and supervised features.

Unsupervised ANNs learn classifications directly from the data they are analyzing and have long been used for image analysis in medicine, including mammography (Floyd Jr. et al., 1994; Leinsinger et al., 2006), MRI (Amatur et al., 1992; Wismüller et al., 2004), and other images (Jiang et al., 2010; Lin et al., 1996). Newer deep learning unsupervised ANNs are still being used for image classification and diagnosis based on images (Kallenberg et al., 2016).

Supervised ANNs require data with known outcomes to help guide the machine learning algorithms used in ANNs. Retrospective medical data is readily available from EHRs and other medical databases to serve as training data for supervised learning methods. Supervised learning ANNs have their neurodes arranged in layers, with increasing numbers of hidden layers enabling a higher degree of nonlinearity in the resultant model (Walczak & Cerpa, 1999).

ANNs have tremendous advantages for performing medical modeling (Tu, 1996; Walczak, 2008). These advantages include: ANN models are created using machine learning so arbitrarily complex non-linear relationships are identifiable; ANNs are nonparametric so prior knowledge of population and error distributions is unnecessary; once trained they are tolerant of noise in the data; and they are extremely fast even when working with very large data sets. Although research with ANNs in medical domains is growing rapidly, the main reason why ANNs have not been implemented and adopted further in clinical settings is their black box nature. The black box nature of ANNs does not enable causal analysis similar to logistic regression (Tu, 1996). Understanding causal relationships and identifying relevant variables is of significant importance in medical decision making (Hunink et al., 2014). However, this should not be seen as an impediment to implementing ANNs in medical decision support systems since various techniques exist to facilitate outcome explanations and assessment of the causal nature between the independent and dependent variables (Andrews et al., 1995; Guidotti et al., 2018; Olden & Jackson, 2002; Walczak, 2008). These techniques include rule extraction by examining the weighted connections (Augusta & Kathirvalavakumar, 2012; Zhou et al., 2003) and variable influence determination by leaving each variable and sets of variables out of the input set and evaluating their relative influence (Tu, 1996; Walczak, 2008).

CURRENT DIRECTIONS OF ANN RESEARCH IN MEDICINE

While research focused on ANNs in medicine is growing rapidly, adoption of ANNs in clinical practice is still lagging. Logistic regression is the principal modeling methodology employed in medical research (Dreiseitl & Ohno-Machado, 2002; Tu, 1996). Most ANN research presents a comparison of ANN prediction results evaluated against results from comparable regression models (e.g., Churpek et al. (2016), Fei et al. (2017), Futoma et al. (2015), Kim et al. (2018), Teo et al. (2018)), and this practice needs to continue for the time being, until most clinicians feel comfortable with ANNs. The ongoing need to compare ANN results against other more familiar statistical models, including regression, is not unique to medical domains. Comparisons of ANN results to other recognized techniques occurs across a wide variety of domains including: engineering (Kaytez et al., 2015; Khademi et al., 2017), geological and environmental sciences (Dou et al., 2018; Nourani et al., 2014), and physics (Kumar et al., 2015) among others. Until the research community accepts the general validity of ANNs, continued presentations of comparative models is beneficial to gaining acceptance of ANN medical models and systems.

In the last decade, the concept of deep learning and Deep ANNs has become a hot topic in machine learning research (Arel et al. 2010). Deep ANNs have proven to be beneficial, especially in sensory recognition and classification problems (Schmidhuber, 2015), such as detecting malignancies in medical images (Hinton, 2018). The definition of a “Deep ANN” is any ANN that has more than one hidden layer. The longer the path between the independent input variables and the dependent output variable, then the deeper the possible learning. One difference between Deep ANNs and more traditional supervised multilayer perceptrons and unsupervised Kohonen self-organizing map (SOM) ANNs is that the individual layers or collections thereof may alternately use different supervised learning or unsupervised learning training methods, thus combining both supervised and unsupervised learning into a single ANN network. It should be noted though that the idea of multiple hidden layer ANNs has been around since at least 1989 (Cybenko, 1989; Funahashi, 1989) or earlier and that the backpropagation learning algorithm created by Werbos (1974), which first enable multilayered ANNs (including the input layer, hidden layer(s), and output layer), does not pose any theoretical limit on the number of hidden layers

within an ANN. While no theoretical limit exists, the addition of each new hidden layer introduces another layer of complexity in the solution surface (e.g., introducing concavities to a convex solution surface) (Walczak & Cerpa, 1999) and in practice most medical research ANN models do not go past 3 hidden layers (Dahl et al., 2012).

From the PubMed data displayed in Figure 1, seven articles published in 2018 had deep learning or Deep ANN in their title (Cao et al., 2018; Gawehn et al., 2018; Grapov et al., 2018; Higaki et al., 2018; Pinaya et al., 2018; Prahs et al., 2018; Zanella-Calzada et al., 2018). A Google Scholar search for research published in 2018 containing the phrase “deep neural network” combined with some of the terms from the PubMed search, specifically: cancer, clinic, hospital, medicine, patient, or surgery; yielded over 17,000 results. From the Google Scholar search, some examples of items with the term deep learning or Deep ANN in their title include: scanning images for cancer markers (Ali et al., 2018; Bevilacqua et al., 2018; Salz et al., 2018), dentistry image scanning (Karimian et al., 2018), and retinal image analysis (Kermamy et al., 2018). These sample articles represent a very small portion of the research currently being conducted on the application of deep learning through Deep ANNs in the field of medicine.

Evolutionary learning is being applied to ANNs more frequently. The term evolutionary learning and also evolutionary ANN indicates that a biological model, which includes artificial life and genetic algorithms, is being applied to an ANN to determine the optimal network architecture or is being used in combination with ANNs to select input criteria or further improve the output classifications or predictions of the ANN. The idea of combining evolutionary learning, in particular genetic algorithms, to improve ANN performance started in the late 1980's (Miller et al., 1989; Montana & Davis, 1989), but has seen significant growth recently including the application of evolutionary learning to Deep ANNs (Arabasadi et al., 2017; Cui et al., 2018; Jiang et al., 2017; Li et al., 2018).

Besides Deep ANNs, a relatively new ANN methodology that is attracting attention for medical image and signal analysis is the convolutional ANN, abbreviated as either CNN or ConvNet. A CNN is composed of multiple layers and after the input layer, alternates between a convolution layer, which tries to match regions of an image or signal, and pooling layers, which condense the output from the convolutional layer into a smaller data set while maintaining all relevant information (Nebauer, 1998). Because of the multiple layers of CNNs, they should also be considered as Deep ANNs. Some CNN configurations specify an additional layer between the convolutional layer and the pooling layer which serves as an activation function to recognize relevant parts of the signal. These layers may be stacked to form multiple convolution to activation to pooling combinations until the signal classification achieves a desired accuracy. Examples of recent application of CNNs in medicine are: detecting myocardial infarction from electrocardiogram signals (Acharya et al., 2017), and detection of possible cancerous lesions for breast (Bejnordi et al., 2017), gastric (Hirasawa et al., 2018), lung (Yu et al., 2017), and pancreatic (Wang et al., 2018) cancers among others (Halicek et al., 2018).

ANNs provide the field of medicine with an efficient and reliable method for performing both non-parametric and nonlinear modeling using machine learning. While resulting classification and prediction models perform very well, a recent trend has emerged to try and further strengthen ANN results and possibly make their outcomes more understandable. This trend is the combination (or triangulation) of ANNs with other mathematical, statistical, and machine learning methods, including other ANNs, into single systems (Walczak, 2012). Smith (1993) indicates that all supervised learning ANN variables should first be evaluated using correlation analysis (e.g., Pearson's correlation matrix) to identify and remove highly correlated variables that can decrease ANN performance accuracy. CNNs, especially if convolution and pooling layers are iterated multiple times, may be considered as a form of triangulation that enables the output of earlier ANN to be used as input to a new ANN, which then improves upon

the prediction or classification of the image from the earlier ANNs. The term ANN ensemble signifies that a combination of ANNs are used with the output of one or more ANNs serving as the input to other ANNs. The idea of utilizing other machine learning methods in combination with ANNs was mentioned earlier in the form of utilizing evolutionary algorithms to improve ANN performance by determining optimal independent input variables or the form of ANN architectures. Traditional statistical methods may be used to help determine the optimal set of independent variables. An example of how to triangulate ANNs with other methods is the classification of genes related to cancers, which uses a statistical method, principal component analysis, to select genes to be considered in the input variable set and an ANN to classify the gene sets related to various cancers based on the variables determined by the principal component analysis statistical method (Chu & Wang, 2009).

ANN triangulation approaches to solving medical problems are varied and use a wide range of techniques in combination with ANNs. Table 1 displays a few recent examples of different techniques combined with ANNs in medical domains, demonstrating the growing trend of combining ANNs with other modeling strategies to gain the nonparametric and machine learning advantages of ANNs while diminishing any disadvantages associated with the black box nature of ANNs.

Table 1. Methods commonly combined (triangulated) with ANNs in medicine.

Triangulated Method	Medical Problem	Reference
Genetic algorithm (evolutionary)	multiple (diabetes, retinopathy) nasopharyngeal carcinoma	(Cortes et al., 2017) (Mohammed et al., 2018)
Mathematical/Statistical: k-means cluster correlation	Heart disease Pancreatic cancer survival	(Malav et al., 2017) (Walczak & Velanovich, 2018)
Other ANN (ensembles)	classifying white blood cells	(Rawat et al., 2018)
Combined: math/stat & ANN ensemble	blood pressure prediction	(Sadrawi et al., 2016)

ANN MEDICAL RESEARCH CONTINUING FORWARD

As stated above, the primary goal of research in medicine should always be to improve the quality of life of patients. This may be achieved through improved diagnostics to better understand and identify medical risks, improved treatment and outcomes assessment to improve recovery and understanding of outcomes, improved workflow and resource efficiencies to decrease the time required to obtain medical consultation or relief, and reduced costs to lessen the financial impact on patients.

Because ANNs utilize machine learning to model complex nonlinear problems, the results may be viewed as heuristic in nature and typically do not achieve 100% accuracy (both sensitivity and specificity). Research is ongoing to try and improve the performance of ANN models in the medical domain to better satisfy the goal of improved patient quality of life. As an example, some of the earliest ANN research was focused on predicting myocardial infarctions (Baxt 1990, 1991, 1992), a type of heart failure. ANN-based research on myocardial infarctions has continued (Baxt et al., 2002) and has sought to utilize new developments in ANN learning methods and architectures such as deep learning implemented in Deep ANNs (Acharya et al., 2017). A timeline of research examining the use of ANNs to detect or diagnose myocardial infarctions is shown in Table 2. We limit the results in Table 2 to 35 articles for readability, though many more exist for this specific medical diagnosis. As ANN researched progresses, in addi-

tion to utilizing newer techniques like convolutional and Deep ANNs, the research has shifted focus to not only diagnose myocardial infarctions, but also to diagnose other related coronary syndromes and to predict outcomes and survival related to myocardial infarctions, encompassing the full diagnostic and prognostic spectrum. As ANN research continues to try and improve on existing diagnostic, prognostic, cost savings, and resource allocation issues in medicine, the trend of expanding the focus of ANN solutions will continue and make ANN-based clinical decision support systems more encompassing.

FUTURE RESEARCH DIRECTIONS

Though ANN solutions in medicine have proven reliable and efficacious, many physicians and other clinical practitioners still remain hesitant to adopt this technology. The reason for this adoption reluctance as mentioned previously, ANNs are considered a black box and the consequent lack of information concerning causal relationships between the independent and dependent variables (Tu, 1996). ANN research should definitely continue and expand to be able to show applicability across all subfields of medicine and further continue research to provide mechanisms to automatically provide knowledge concerning the casual effects discovered by ANNs, since this knowledge is critical to medical research. Some ANN development tools, like JustNN (Neural Planner Software, 2015) imply variable causality by summing the absolute weights of connections.

Table 2. ANN research to diagnose acute myocardial infarction (AMI).

YEAR	# of articles	References
1990	1	(Baxt, 1990)
1991	2	(Baxt, 1991); (Furlong et al., 1991)
1992	2	(Baxt, 1992); (Reddy et al., 1992)
1993	1	(Bortolan & Willems, 1993)
1994	3	(Baxt, 1994); (Hedén et al., 1994); (Yang et al., 1994)
1995	2	(Baxt & White, 1995); (Selker et al., 1995)
1996	4	(Baxt & Skora, 1996); (Hedén et al., 1996); (Kumaravel et al., 1996); (Pedersen et al., 1996)
1997	3	(Ellenius et al., 1997); (Hedén et al., 1997); (Kennedy et al., 1997)
2000	1	(Ohlsson et al., 2000)
2001	1	(Ohlsson et al., 2001)
2002	2	(Baxt et al., 2002); (Olsson et al., 2002)
2004	2	(Haraldsson et al., 2004); (Hollander et al., 2004)
2005	2	(Bigi et al., 2005) [†] ; (Harrison & Kennedy, 2005) ^{††}
2007	1	(Eggers et al., 2007) ^{†††}
2010	1	(Arif et al., 2010)
2011	1	(Sankari & Adeli, 2011) ^{††}
2013	1	(Keshtkar et al., 2013)
2014	1	(Safdarian et al., 2014)
2015	1	(Bhaskar, 2015)
2017	1	(Acharya et al., 2017)
2018	3	(Liu et al., 2018a); (Liu et al., 2018b, 2018c)

[†]examines recovery/prognostics following AMI

^{††}models a group of acute coronary syndromes including myocardial infarction

^{†††}also models infarct size

The expansion of ANN research to ultimately cover all fields of medicine is a desirable goal. One step in this direction that is needed is to make ANN solutions more generalizable. As an example, ANN research on predicting transfusion requirements has uniformly been targeted at predicting transfusion needs for patients of a single medical illness or procedure. Examples include ANN networks to predict transfusion requirements for coronary artery bypass (CABG) surgery (Covin et al., 2003), leukemia patients (Ho & Chang, 2011), and abdominal aortic aneurysm (AAA) surgery (Walczak & Scharf, 2000). Future research is needed to identify, using ANNs, those variables causally related to transfusion needs in humans and then develop ANN models that may be applied across a much wider range of surgery types and illnesses that may require transfusions. Ideally, new ANN research will produce a combined resource and prognostic model that is applicable to all classes of surgery. This approach of generalization should also be applied to other prognostic problems, including the anesthesia, physical and occupational therapy, and pharmacological needs of patients.

Electronic health records (EHRs) are widely used worldwide to improve patient care (Bourgeois & Yaylacicegi, 2010; Mennemeyer et al., 2016) and also serve as a data resource for ANN research (Bibault et al., 2016; Weng et al., 2017). As predicted by Kuperman et al. (2007) ANN diagnostic and prognostic tools need to become integrated into EHRs to provide decision support. ANNs are starting to be embedded into the clinical decision making process, especially in radiology (Pesapane et al., 2018). The integration of ANN diagnostic or prognostic models into EHRs, with a corresponding explanation facility, will provide another route for implementation of ANNs in practice. The need for moving ANN research into clinical practice is widely recognized (Khazaei et al., 2018).

The future of medicine is moving towards body area networks and implantable medical devices (IMD) to enable mobile data acquisition and treatment decision making. ANNs are already starting to be embedded within physical therapy wireless body area network devices (Lin et al., 2015), and this trend of utilizing ANNs in body area networks and IMDs will continue to grow. The integration of ANNs into an IMD requires the overall size of the implemented ANN to be small, but this is another future direction for ANN research.

CONCLUSION

ANNs are a machine learning method based on the neurology of the human brain and are useful for classification and prediction problems in medicine. Past research has demonstrated the efficacy of utilizing ANNs to perform diagnostics, prognostics, cost reduction, and resource allocation problem solving in medicine. The use of ANNs to model medical diagnostic, prognostic, resource allocation, and cost minimization problems is a growing trend for the foreseeable future.

REFERENCES

- Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences*, 415, 190–198. doi:10.1016/j.ins.2017.06.027
- Ali, I., Hart, G., Gunabushanam, G., Liang, Y., Muhammad, W., Nartowt, B., Kane, M., Ma, X., & Deng, J. (2018). Lung nodule detection via deep reinforcement learning. *Frontiers in Oncology*, 8, 108. doi:10.3389/fonc.2018.00108 PMID:29713615

Amartur, S. C., Piraino, D., & Takefuji, Y. (1992). Optimization neural networks for the segmentation of magnetic resonance images. *IEEE Transactions on Medical Imaging*, 11(2), 215–220. doi:10.1109/42.141645 PMID:18218375

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6), 373–389. doi:10.1016/0950-7051(96)81920-4

Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Computer Methods and Programs in Biomedicine*, 141, 19–26. doi:10.1016/j.cmpb.2017.01.004 PMID:28241964

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine*, 5(4), 13–18. doi:10.1109/MCI.2010.938364

Arif, M., Malagore, I. A., & Afsar, F. A. (2010). Automatic detection and localization of myocardial infarction using back propagation neural networks. In *4th International Conference on Bioinformatics and Biomedical Engineering*, (pp. 1-4). IEEE. 10.1109/ICBBE.2010.5514664

Asada, N., Doi, K., MacMahon, H., Montner, S. M., Giger, M. L., Abe, C., & Wu, Y. U. Z. H. E. N. G. (1990). Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: Pilot study. *Radiology*, 177(3), 857–860. doi:10.1148/radiology.177.3.2244001 PMID:2244001

Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. *Neural Processing Letters*, 35(2), 131–150. doi:10.1007/11063-011-9207-8

Baxt, W. G. (1990). Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion. *Neural Computation*, 2(4), 480–489. doi:10.1162/neco.1990.2.4.480

Baxt, W. G. (1991). Use of an artificial neural network for the diagnosis of myocardial infarction. *Annals of Internal Medicine*, 115(11), 843–848. doi:10.7326/0003-4819-115-11-843 PMID:1952470

Baxt, W. G. (1992). Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction. *Annals of Emergency Medicine*, 21(12), 1439–1444. doi:10.1016/S0196-0644(05)80056-3 PMID:1443838

Baxt, W. G. (1994). A neural network trained to identify the presence of myocardial infarction bases some decisions on clinical associations that differ from accepted clinical teaching. *Medical Decision Making*, 14(3), 217–222. doi:10.1177/0272989X9401400303 PMID:7934708

Baxt, W. G., Shofer, F. S., Sites, F. D., & Hollander, J. E. (2002). A neural computational aid to the diagnosis of acute myocardial infarction. *Annals of Emergency Medicine*, 39(4), 366–373. doi:10.1067/mem.2002.122705 PMID:11919522

Baxt, W. G., & Skora, J. (1996). Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet*, 347(8993), 12–14. doi:10.1016/S0140-6736(96)91555-X PMID:8531540

Baxt, W. G., & White, H. (1995). Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction. *Neural Computation*, 7(3), 624–638. doi:10.1162/neco.1995.7.3.624 PMID:8935962

- Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karsssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermesen, M., Manson, Q. F., Balkenhol, M., & Geessink, O. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Journal of the American Medical Association*, 318(22), 2199–2210. doi:10.1001/jama.2017.14585 PMID:29234806
- Bevilacqua, V., Brunetti, A., Guerriero, A., Trotta, G. F., Telegrafo, M., & Moschetta, M. (2018). A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cognitive Systems Research*.
- Bhaskar, N. A. (2015). Performance analysis of support vector machine and neural networks in detection of myocardial infarction. *Procedia Computer Science*, 46, 20–30. doi:10.1016/j.procs.2015.01.043
- Bibault, J. E., Giraud, P., & Burgun, A. (2016). Big data and machine learning in radiation oncology: State of the art and future prospects. *Cancer Letters*, 382(1), 110–117. doi:10.1016/j.canlet.2016.05.033 PMID:27241666
- Bigi, R., Gregori, D., Cortigiani, L., Desideri, A., Chiarotto, F. A., & Toffolo, G. M. (2005). Artificial neural networks and robust Bayesian classifiers for risk stratification following uncomplicated myocardial infarction. *International Journal of Cardiology*, 101(3), 481–487. doi:10.1016/j.ijcard.2004.07.008 PMID:15907418
- Bortolan, G., & Willems, J. L. (1993). Diagnostic ECG classification based on neural networks. *Journal of Electrocardiology*, 26, 75–79. PMID:8189152
- Bourgeois, S., & Yaylacicegi, U. (2010). Electronic Health Records: Improving Patient Safety and Quality of Care in Texas Acute Care Hospitals. *International Journal of Healthcare Information Systems and Informatics*, 5(3), 1–13. doi:10.4018/jhisi.2010070101
- Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., & Xie, Z. (2018). Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics & Bioinformatics*, 16(1), 17–32. doi:10.1016/j.gpb.2017.07.003 PMID:29522900
- Chu, F., & Wang, L. (2009). Biomedical Data Mining Using RBF Neural Networks. In *Medical Informatics: Concepts, Methodologies, Tools, and Applications* (pp. 2066–2073). IGI Global. doi:10.4018/978-1-60566-050-9.ch157
- Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2), 368–374. doi:10.1097/CCM.0000000000001571 PMID:26771782
- Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., & Yang, S. (2017). AdaNet: Adaptive Structural Learning of Artificial Neural Networks. In *International Conference on Machine Learning* (pp. 874–883).
- Covin, R., O'Brien, M., Grunwald, G., Brimhall, B., Sethi, G., Walczak, S., Reiquam, W., Rajagopalan, C., & Shroyer, A. L. (2003). Factors affecting transfusion of fresh frozen plasma, platelets, and red blood cells during elective coronary artery bypass graft surgery. *Archives of Pathology & Laboratory Medicine*, 127(4), 415–423. PMID:12683868

Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An Improved Support Vector Machine-based Diabetic Readmission Prediction. *Computer Methods and Programs in Biomedicine*, 166, 123–135. doi:10.1016/j.cmpb.2018.10.012 PMID:30415712

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. doi:10.1007/BF02551274

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 30–42. doi:10.1109/TASL.2011.2134090

Dassen, W. R., Mulleneers, R. G., Dulk, K. D., Smeets, J. R., Cruz, F., Penn, O. C., & Wellens, H. J. (1990). An Artificial Neural Network to Localize Atrioventricular Accessory Pathways in Patients Suffering from the Wolff-Parkinson-White Syndrome. *Pacing and Clinical Electrophysiology*, 13(12), 1792–1796. doi:10.1111/j.1540-8159.1990.tb06892.x PMID:1704543

Dou, J., Yamagishi, H., Zhu, Z., Yunus, A. P., & Chen, C. W. (2018). TXT-tool 1.081-6.1 A Comparative Study of the Binary Logistic Regression (BLR) and Artificial Neural Network (ANN) Models for GIS-Based Spatial Predicting Landslides at a Regional Scale. In *Landslide Dynamics: ISDR-ICL Landslide Interactive Teaching Tools* (pp. 139–151). Springer. doi:10.1007/978-3-319-57774-6_10

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6), 352–359. doi:10.1016/S1532-0464(03)00034-0 PMID:12968784

Dybowski, R. (2000). Neural computation in medicine: perspectives and prospects. In *Artificial Neural Networks in Medicine and Biology* (pp. 26–36). Springer. doi:10.1007/978-1-4471-0513-8_4

Eggers, K. M., Ellenius, J., Dellborg, M., Groth, T., Oldgren, J., Swahn, E., & Lindahl, B. (2007). Artificial neural network algorithms for early diagnosis of acute myocardial infarction and prediction of infarct size in chest pain patients. *International Journal of Cardiology*, 114(3), 366–374. doi:10.1016/j.ijcard.2005.12.019 PMID:16797088

Ellenius, J., Groth, T., Lindahl, B., & Wallentin, L. (1997). Early assessment of patients with suspected acute myocardial infarction by biochemical monitoring and neural network analysis. *Clinical Chemistry*, 43(10), 1919–1925. doi:10.1093/clinchem/43.10.1919 PMID:9342013

Fei, Y., Hu, J., Gao, K., Tu, J., Li, W. Q., & Wang, W. (2017). Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models. *Journal of Critical Care*, 39, 115–123. doi:10.1016/j.jcrc.2017.02.032 PMID:28246056

Floyd, C. E. Jr, Lo, J. Y., Yun, A. J., Sullivan, D. C., & Kornguth, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 74(11), 2944–2948. doi:10.1002/1097-0142(19941201)74:11<2944::AID-CNCR2820741109>3.0.CO;2-F PMID:7954258

Funahashi, K. I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3), 183–192. doi:10.1016/0893-6080(89)90003-8

- Furlong, J. W., Dupuy, M. E., & Heinsimer, J. A. (1991). Neural Network Analysis of Serial Cardiac Enzyme Data A Clinical Application of Artificial Machine Intelligence. *American Journal of Clinical Pathology*, 96(1), 134–141. doi:10.1093/ajcp/96.1.134 PMID:2069131
- Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56, 229–238. doi:10.1016/j.jbi.2015.05.016 PMID:26044081
- Gawehn, E., Hiss, J. A., Brown, J. B., & Schneider, G. (2018). Advancing drug discovery via GPU-based deep learning. *Expert Opinion on Drug Discovery*, 13(7), 579–582. doi:10.1080/17460441.2018.1465407 PMID:29668343
- Grapov, D., Fahrman, J., Wanichthanarak, K., & Khoomrung, S. (2018). Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. *OMICS: A Journal of Integrative Biology*, 22(10), 22. doi:10.1089/omi.2018.0097 PMID:30124358
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 93.
- Halicek, M., Little, J. V., Wang, X., Patel, M., Griffith, C. C., El-Deiry, M. W., Chen, A. Y., & Fei, B. (2018). Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks. In *Optical Imaging, Therapeutics, and Advanced Technology in Head and Neck Surgery and Otolaryngology 2018* (Vol. 10469, p. 104690X). International Society for Optics and Photonics. doi:10.1117/12.2289023
- Haraldsson, H., Edenbrandt, L., & Ohlsson, M. (2004). Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks. *Artificial Intelligence in Medicine*, 32(2), 127–136. doi:10.1016/j.artmed.2004.01.003 PMID:15364096
- Harrison, R. F., & Kennedy, R. L. (2005). Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Annals of Emergency Medicine*, 46(5), 431–439. doi:10.1016/j.annemergmed.2004.09.012 PMID:16271675
- Hedén, B., Edenbrandt, L., Haisty, W. K. Jr., & Pahlm, O. (1994). Artificial neural networks for the electrocardiographic diagnosis of healed myocardial infarction. *The American Journal of Cardiology*, 74(1), 5–8. doi:10.1016/0002-9149(94)90482-0 PMID:8017306
- Hedén, B., Ohlin, H., Rittner, R., & Edenbrandt, L. (1997). Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation*, 96(6), 1798–1802. doi:10.1161/01.CIR.96.6.1798 PMID:9323064
- Hedén, B. O., Ohlsson, M., Rittner, R., Pahlm, O., Haisty, W. K., Peterson, C., & Edenbrandt, L. (1996). Agreement between artificial neural networks and experienced electrocardiographer on electrocardiographic diagnosis of healed myocardial infarction. *Journal of the American College of Cardiology*, 28(4), 1012–1016. PMID:8837583
- Higaki, A., Mogi, M., Iwanami, J., Min, L. J., Bai, H. Y., Shan, B. S., Kukida, M., Kan-no, H., Ikeda, S., Higaki, J., & Horiuchi, M. (2018). Predicting outcome of Morris water maze test in vascular dementia mouse model with deep learning. *PLoS One*, 13(2), e0191708. doi:10.1371/journal.pone.0191708 PMID:29415035

- Hinton, G. (2018). Deep learning—A technology with the potential to transform health care. *Journal of the American Medical Association*, 320(11), 1101–1102. doi:10.1001/jama.2018.11100 PMID:30178065
- Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., Ohnishi, T., Fujishiro, M., Matsuo, K., Fujisaki, J., & Tada, T. (2018). Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer*, 21(4), 653–660. doi:10.1007/10120-018-0793-2 PMID:29335825
- Ho, W. H., & Chang, C. S. (2011). Genetic-algorithm-based artificial neural network modeling for platelet transfusion requirements on acute myeloblastic leukemia patients. *Expert Systems with Applications*, 38(5), 6319–6323. doi:10.1016/j.eswa.2010.11.110
- Hollander, J. E., Sease, K. L., Sparano, D. M., Sites, F. D., Shofer, F. S., & Baxt, W. G. (2004). Effects of neural network feedback to physicians on admit/discharge decision for emergency department patients with chest pain. *Annals of Emergency Medicine*, 44(3), 199–205. doi:10.1016/j.annemergmed.2004.02.037 PMID:15332058
- Hunink, M. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., & Glasziou, P. P. (2014). *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge University Press. doi:10.1017/CBO9781139506779
- Jiang, J., Trundle, P., & Ren, J. (2010). Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8), 617–631. doi:10.1016/j.compmedimag.2010.07.003 PMID:20713305
- Jiang, S., Chin, K. S., Wang, L., Qu, G., & Tsui, K. L. (2017). Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. *Expert Systems with Applications*, 82, 216–230. doi:10.1016/j.eswa.2017.04.017
- Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., Vachon, C. M., Holland, K., Winkel, R. R., & Lillholm, M. (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, 35(5), 1322–1331. doi:10.1109/TMI.2016.2532122 PMID:26915120
- Karimian, N., Salehi, H. S., Mahdian, M., Alnajjar, H., & Tadinada, A. (2018). Deep learning classifier with optical coherence tomography images for early dental caries detection. In *Lasers in Dentistry XXIV* (Vol. 10473, p. 1047304). International Society for Optics and Photonics.
- Kaytez, F., Taplamacioglu, M. C., Cam, E., & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems*, 67, 431–438. doi:10.1016/j.ijepes.2014.12.036
- Kennedy, R. L., Harrison, R. F., Burton, A. M., Fraser, H. S., Hamer, W. G., MacArthur, D., McAllum, R., & Steedman, D. J. (1997). An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: Evaluation and comparison with serum myoglobin measurements. *Computer Methods and Programs in Biomedicine*, 52(2), 93–103. doi:10.1016/S0169-2607(96)01782-8 PMID:9034674

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131. doi:10.1016/j.cell.2018.02.010 PMID:29474911

Keshtkar, A., Seyedarabi, H., Sheikhzadeh, P., & Rasta, S. H. (2013). Discriminant analysis between myocardial infarction patients and healthy subjects using Wavelet Transformed signal averaged electrocardiogram and probabilistic neural network. *Journal of Medical Signals and Sensors*, 3(4), 225. doi:10.4103/2228-7477.128316 PMID:24696156

Khademi, F., Akbari, M., Jamal, S. M., & Nikoo, M. (2017). Multiple linear regression, artificial neural network, and fuzzy logic prediction of 28 days compressive strength of concrete. *Frontiers of Structural and Civil Engineering*, 11(1), 90–99. doi:10.1007/11709-016-0363-9

Khazaei, P. R., Bagherzadeh, J., Niazkhani, Z., & Pirnejad, H. (2018). A dynamic model for predicting graft function in kidney recipients' upcoming follow up visits: A clinical application of artificial neural network. *International Journal of Medical Informatics*, 119, 125–133. doi:10.1016/j.ijmedinf.2018.09.012 PMID:30342680

Kim, J. S., Merrill, R. K., Arvind, V., Kaji, D., Pasik, S. D., Nwachukwu, C. C., Vargas, L., Osman, N. S., Oermann, E. K., Caridi, J. M., & Cho, S. K. (2018). Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. *Spine*, 43(12), 853–860. doi:10.1097/BRS.0000000000002442 PMID:29016439

Kumar, R., Aggarwal, R. K., & Sharma, J. D. (2015). Comparison of regression and artificial neural network models for estimation of global solar radiations. *Renewable & Sustainable Energy Reviews*, 52, 1294–1299. doi:10.1016/j.rser.2015.08.021

Kumaravel, N., Sridhar, K. S., & Nithiyanandam, N. (1996). Automatic diagnosis of heart diseases using neural network. In *Proceedings of the 1996 Fifteenth Southern Biomedical Engineering Conference*, (pp. 319-322). IEEE. 10.1109/SBEC.1996.493214

Kuperman, G. J., Bobb, A., Payne, T. H., Avery, A. J., Gandhi, T. K., Burns, G., Classen, D. C., & Bates, D. W. (2007). Medication-related clinical decision support in computerized provider order entry systems: A review. *Journal of the American Medical Informatics Association*, 14(1), 29–40. doi:10.1197/jamia.M2170 PMID:17068355

Leinsinger, G., Schlossbauer, T., Scherr, M., Lange, O., Reiser, M., & Wismüller, A. (2006). Cluster analysis of signal-intensity time course in dynamic breast MRI: Does unsupervised vector quantization help to evaluate small mammographic lesions? *European Radiology*, 16(5), 1138–1146. doi:10.1007/00330-005-0053-9 PMID:16418862

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident; Analysis and Prevention*, 111, 354–363. doi:10.1016/j.aap.2017.11.028 PMID:29276978

Lin, H. C., Chiang, S. Y., Lee, K., & Kan, Y. C. (2015). An activity recognition model using inertial sensor nodes in a wireless sensor network for frozen shoulder rehabilitation exercises. *Sensors (Basel)*, 15(1), 2181–2204. doi:10.3390/150102181 PMID:25608218

- Lin, J. S., Cheng, K. S., & Mao, C. W. (1996). A fuzzy Hopfield neural network for medical image segmentation. *IEEE Transactions on Nuclear Science*, 43(4), 2389–2398. doi:10.1109/23.531787
- Liu, N., Wang, L., Chang, Q., Xing, Y., & Zhou, X. (2018a). A Simple and Effective Method for Detecting Myocardial Infarction Based on Deep Convolutional Neural Network. *Journal of Medical Imaging and Health Informatics*, 8(7), 1508–1512. doi:10.1166/jmihi.2018.2463
- Liu, W., Huang, Q., Chang, S., Wang, H., & He, J. (2018b). Multiple-feature-branch convolutional neural network for myocardial infarction diagnosis using electrocardiogram. *Biomedical Signal Processing and Control*, 45, 22–32. doi:10.1016/j.bspc.2018.05.013
- Liu, W., Zhang, M., Zhang, Y., Liao, Y., Huang, Q., Chang, S., Wang, H., & He, J. (2018c). Real-time multilead convolutional neural network for myocardial infarction detection. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1434–1444. doi:10.1109/JBHI.2017.2771768 PMID:29990164
- Malav, A., Kadam, K., & Kamat, P. (2017). Prediction of Heart Disease Using K-Means and Artificial Neural Network as Hybrid Approach to Improve Accuracy. *IACSIT International Journal of Engineering and Technology*, 9(4), 3081–3085. doi:10.21817/ijet/2017/v9i4/170904101
- Mennemeyer, S. T., Menachemi, N., Rahurkar, S., & Ford, E. W. (2016). Impact of the HITECH act on physicians' adoption of electronic health records. *Journal of the American Medical Informatics Association*, 23(2), 375–379. doi:10.1093/jamia/ocv103 PMID:26228764
- Miller, G. F., Todd, P. M., & Hegde, S. U. (1989). Designing Neural Networks using Genetic Algorithms. In *Proceedings of the Third International Conference on Genetic Algorithms* (pp. 379–384).
- Miller, R. A. (1994). Medical diagnostic decision support systems—past, present, and future: A threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, 1(1), 8–27. doi:10.1136/jamia.1994.95236141 PMID:7719792
- Mohammed, M. A., Ghani, M. K. A., Arunkumar, N., Hamed, R. I., Abdullah, M. K., & Burhanuddin, M. A. (2018). A real time computer aided object detection of nasopharyngeal carcinoma using genetic algorithm and artificial neural network based on Haar feature fear. *Future Generation Computer Systems*, 89, 539–547. doi:10.1016/j.future.2018.07.022
- Montana, D. J., & Davis, L. (1989). Training Feedforward Neural Networks Using Genetic Algorithms. In *International Joint Conference on Artificial Intelligence* (Vol. 89, pp. 762–767).
- Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1. doi:10.1186/40537-014-0007-7
- Nebauer, C. (1998). Evaluation of convolutional neural networks for visual recognition. *IEEE Transactions on Neural Networks*, 9(4), 685–696. doi:10.1109/72.701181 PMID:18252491
- Neural Planner Software. (2015). *JustNN Help User Guide*. Retrieved from: <http://www.justnn.com/108491/JustNN.pdf>
- Nourani, V., Pradhan, B., Ghaffari, H., & Sharifi, S. S. (2014). Landslide susceptibility mapping at Zonouz Plain, Iran using genetic programming and comparison with frequency ratio, logistic regression, and artificial neural network models. *Natural Hazards*, 71(1), 523–547. doi:10.1007/11069-013-0932-3

- Ohlsson, M., Holst, H., & Edenbrandt, L. (2000). Acute myocardial infarction: analysis of the ECG using artificial neural networks. In *Artificial Neural Networks in Medicine and Biology* (pp. 209–214). Springer. doi:10.1007/978-1-4471-0513-8_31
- Ohlsson, M., Öhlin, H., Wallerstedt, S. M., & Edenbrandt, L. (2001). Usefulness of serial electrocardiograms for diagnosis of acute myocardial infarction. *The American Journal of Cardiology*, 88(5), 478–481. doi:10.1016/S0002-9149(01)01722-2 PMID:11524053
- Olden, J. D., & Jackson, D. A. (2002). Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1-2), 135–150. doi:10.1016/S0304-3800(02)00064-9
- Olsson, S. E., Ohlsson, M., Öhlin, H., & Edenbrandt, L. (2002). Neural networks—a diagnostic tool in acute myocardial infarction with concomitant left bundle branch block. *Clinical Physiology and Functional Imaging*, 22(4), 295–299. doi:10.1046/j.1475-097X.2002.00433.x PMID:12402453
- Pastur-Romay, L. A., Cedrón, F., Pazos, A., & Porto-Pazos, A. B. (2016). Deep artificial neural networks and neuromorphic chips for big data analysis: Pharmaceutical and bioinformatics applications. *International Journal of Molecular Sciences*, 17(8), 1313. doi:10.3390/ijms17081313 PMID:27529225
- Pedersen, S. M., Jørgensen, J. S., & Pedersen, J. B. (1996). Use of neural networks to diagnose acute myocardial infarction. II. A clinical application. *Clinical Chemistry*, 42(4), 613–617. doi:10.1093/clinchem/42.4.613 PMID:8605680
- Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. *Insights Into Imaging*, 9(5), 1–9. doi:10.1007/13244-018-0645-y PMID:30112675
- Pinaya, W. H., Mechelli, A., & Sato, J. R. (2018). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human Brain Mapping*. PMID:30311316
- Prahs, P., Radeck, V., Mayer, C., Cvetkov, Y., Cvetkova, N., Helbig, H., & Märker, D. (2018). OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications. *Graefes Archive for Clinical and Experimental Ophthalmology*, 256(1), 91–98. doi:10.1007/00417-017-3839-y PMID:29127485
- Rawat, J., Singh, A., Bhadauria, H. S., Virmani, J., & Devgun, J. S. (2018). Application of ensemble artificial neural network for the classification of white blood cells using microscopic blood images. *International Journal of Computational Systems Engineering*, 4(2-3), 202–216. doi:10.1504/IJCSYSE.2018.091407
- Reddy, M. R. S., Edenbrandt, L., Svensson, J., Haisty, W. K., & Pahlm, O. (1992). Neural network versus electrocardiographer and conventional computer criteria in diagnosing anterior infarct from the ECG. [IEEE.]. *Proceedings of Computers in Cardiology*, 1992, 667–670. doi:10.1109/CIC.1992.269345
- Sadrawi, M., Shieh, J. S., Fan, S. Z., Lin, C. H., Haraikawa, K., Chien, J. C., & Abbod, M. F. (2016). Intermittent blood pressure prediction via multiscale entropy and ensemble artificial neural networks. In *Biomedical Engineering and Sciences (IECBES), 2016 IEEE EMBS Conference on* (pp. 356–359). IEEE. 10.1109/IECBES.2016.7843473

Safdarian, N., Dabanloo, N. J., & Attarodi, G. (2014). A new pattern recognition method for detection and localization of myocardial infarction using T-wave integral and total integral as extracted features from one cycle of ECG signal. *Journal of Biomedical Science and Engineering*, 7(10), 818–824. doi:10.4236/jbise.2014.710081

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K. R., Zhao, T., Batiste, R., Van Arnam, J., Shmulevich, I., Rao, A. U. K., Lazar, A. J., Sharma, A., Thorsson, V., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., ... Mariamidze, A. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports*, 23(1), 181–193. doi:10.1016/j.celrep.2018.03.086 PMID:29617659

Sankari, Z., & Adeli, H. (2011). HeartSaver: A mobile cardiac monitoring system for auto-detection of atrial fibrillation, myocardial infarction, and atrio-ventricular block. *Computers in Biology and Medicine*, 41(4), 211–220. doi:10.1016/j.compbiomed.2011.02.002 PMID:21377149

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. doi:10.1016/j.neunet.2014.09.003 PMID:25462637

Selker, H. P., Griffith, J. L., Patil, S., Long, W. J., & d'Agostino, R. B. (1995). A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine: The Official Publication of the American Federation for Clinical Research*, 43(5), 468–476. PMID:8528758

Smith, M. (1993). *Neural Networks for Statistical Modeling*. Van Nostrand Reinhold.

Teo, T. P., Ahmed, S. B., Kawalec, P., Alayoubi, N., Bruce, N., Lyn, E., & Pistorius, S. (2018). Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories. *Medical Physics*, 45(2), 830–845. doi:10.1002/mp.12731 PMID:29244902

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11), 1225–1231. doi:10.1016/S0895-4356(96)00002-9 PMID:8892489

Wager, K. A., Lee, F. W., & Glaser, J. P. (2013). *Health Care Information Systems: A Practical Approach for Health Care Management*. Jossey-Bass/Wiley.

Walczak, S. (2008). Evaluating medical decision making heuristics and other business heuristics with neural networks. In *Intelligent Decision Making: An AI-Based Approach* (pp. 259–287). Springer. doi:10.1007/978-3-540-76829-6_10

Walczak, S. (2012). Methodological triangulation using neural networks for business research. *Advances in Artificial Neural Systems*, 2012, 1–12. doi:10.1155/2012/517234

Walczak, S., & Cerpa, N. (1999). Heuristic principles for the design of artificial neural networks. *Information and Software Technology*, 41(2), 107–117. doi:10.1016/S0950-5849(98)00116-5

Walczak, S., & Scharf, J. E. (2000). Reducing surgical patient costs through use of an artificial neural network to predict transfusion requirements. *Decision Support Systems*, 30(2), 125–138. doi:10.1016/S0167-9236(00)00093-2

Walczak, S., & Velanovich, V. (2018). Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks. *Decision Support Systems*, 106, 110–118. doi:10.1016/j.dss.2017.12.007

Wang, S., Wang, R., Zhang, S., Li, R., Fu, Y., Sun, X., Li, Y., Sun, X., Jiang, X., Guo, X., Zhou, X., Chang, J., & Peng, W. (2018). 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as groundglass nodules with diameters ≤ 3 cm using HRCT. *Quantitative Imaging in Medicine and Surgery*, 8(5), 491–499. doi:10.21037/qims.2018.06.03 PMID:30050783

Wang, Y., Wang, D., Ye, X., Wang, Y., Yin, Y., & Jin, Y. (2019). A Tree Ensemble-Based Two-Stage Model for Advanced-Stage Colorectal Cancer Survival Prediction. *Information Sciences*, 474, 106–124. doi:10.1016/j.ins.2018.09.046

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*, 12(4), e0174944. doi:10.1371/journal.pone.0174944 PMID:28376093

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences* (Doctoral Dissertation). Harvard University, Boston, MA.

Wismüller, A., Meyer-Bäse, A., Lange, O., Auer, D., Reiser, M. F., & Sumners, D. (2004). Model-free functional MRI analysis based on unsupervised clustering. *Journal of Biomedical Informatics*, 37(1), 10–18. doi:10.1016/j.jbi.2003.12.002 PMID:15016382

Yang, T. F., Devine, B., & Macfarlane, P. W. (1994). Use of artificial neural networks within deterministic logic for the computer ECG diagnosis of inferior myocardial infarction. *Journal of Electrocardiology*, 27, 188–193. doi:10.1016/S0022-0736(94)80090-1 PMID:7884359

Yu, D., Zhou, M., Yang, F., Dong, D., Gevaert, O., Liu, Z., Shi, J., & Tian, J. (2017). Convolutional neural networks for predicting molecular profiles of non-small cell lung cancer. In *Biomedical Imaging (ISBI 2017)*, 2017 IEEE 14th International Symposium on (pp. 569–572). IEEE. 10.1109/ISBI.2017.7950585

Zanella-Calzada, L., Galván-Tejada, C., Chávez-Lamas, N., Rivas-Gutierrez, J., Magallanes-Quintanar, R., Celaya-Padilla, J., Galván-Tejada, J., & Gamboa-Rosales, H. (2018). Deep Artificial Neural Networks for the Diagnostic of Caries Using Socioeconomic and Nutritional Features as Determinants: Data from NHANES 2013–2014. *Bioengineering (Basel, Switzerland)*, 5(2), 47. doi:10.3390/bioengineering5020047 PMID:29912173

Zhou, Z. H., Jiang, Y., & Chen, S. F. (2003). Extracting symbolic rules from trained neural network ensembles. *AI Communications*, 16(1), 3–15.

ADDITIONAL READING

DeTienne, K. B., DeTienne, D. H., & Joshi, S. A. (2003). Neural Networks as Statistical Tools for Business Researchers. *Organizational Research Methods*, 6(2), 236–265. doi:10.1177/1094428103251907

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial Neural Networks: A Tutorial. *Computer*, 29(3), 31–44. doi:10.1109/2.485891

Penny, W., & Frost, D. (1996). Neural networks in clinical medicine. *Medical Decision Making*, 16(4), 386–398. doi:10.1177/0272989X9601600409 PMID:8912300

Reggia, J. A. (1993). Neural computation in medicine. *Artificial Intelligence in Medicine*, 5(2), 143–157. doi:10.1016/0933-3657(93)90014-T PMID:8358491

Rodvold, D. M., McLeod, D. G., Brandt, J. M., Snow, P. B., & Murphy, G. P. (2001). Introduction to artificial neural networks for physicians: Taking the lid off the black box. *The Prostate*, 46(1), 39–44. doi:10.1002/1097-0045(200101)46:1<39::AID-PROS1006>3.0.CO;2-M PMID:11170130

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The American Statistician*, 50(4), 284–293.

Whitley, D. (1995). Genetic algorithms and neural networks. *Genetic Algorithms in Engineering and Computer Science*, 3, 203–216.

KEY TERMS AND DEFINITIONS

Big Data: Is a term referring to an extremely large collection of data, typically in the petabyte range or larger. The data is only considered “big” if it taxes current state-of-the-art computing capabilities.

Diagnostic: Referring to the diagnosis or determination of a disease or diseases or trauma; defining the medical condition of a patient.

Kohonen/SOM: Is a type of ANN that uses unsupervised learning. Commonly used in audio and visual perception ANNs.

Perceptron/Multilayer Perceptron: Perceptron was an early name for a single processing element, now more commonly called a neuron or neurode. Multilayer perceptrons (MLP) are arrangements of perceptrons into layers, such as the input layer connected to a hidden layer which is then connected to an output layer. The abbreviation MLP is often used to refer to backpropagation trained supervised learning ANNs with at least one hidden layer.

Prognostic: Referring to the treatment and follow-up plan for a specific medical condition or type of trauma, for example, a course of chemo-therapy and radiation prior to surgery for the reduction and removal of a cancerous fibroid.

Sensitivity: The percentage of patients with a specific diagnosis or outcome that are correctly identified, which is the number of true positives divided by the sum of the true positives and false negatives.

Specificity: The percentage of patients who do not have a specific diagnosis or desired outcome that are correctly identified, which is the number of true negatives divided by the sum of the true negatives and false positives.

Internet of Things and Data Science in Healthcare

16

George Tzanis*Aristotle University of Thessaloniki, Greece***Ourania-Ioanna Fotopoulou***Aristotle University of Thessaloniki, Greece*

INTRODUCTION

Health care advances are taking place rapidly all over the world. The newest medical discoveries and the incorporation of best available evidence in medical practice have guided to serious improvements in human health and well-being. Life expectancy has been considerably increasing, leading to an aging population with different needs and desires. Moreover, world population has been rising exponentially, challenging the effectiveness of healthcare systems. New predictive tests, advanced procedures, and genomic therapy medicines have made medical care increasingly expensive, forcing cost containment even while overall effectiveness improves (Vasanth & Sbert, 2012). Nowadays, prevention and early detection of disease signs as well as quality of life are becoming the points of focus more than ever before. Furthermore, patient-centered medicine is deemed particularly important and promises better disease treatment for each individual. Patients are becoming better-informed, want to know more about their health and their treatment, and demand to be involved in decision-making regarding their health.

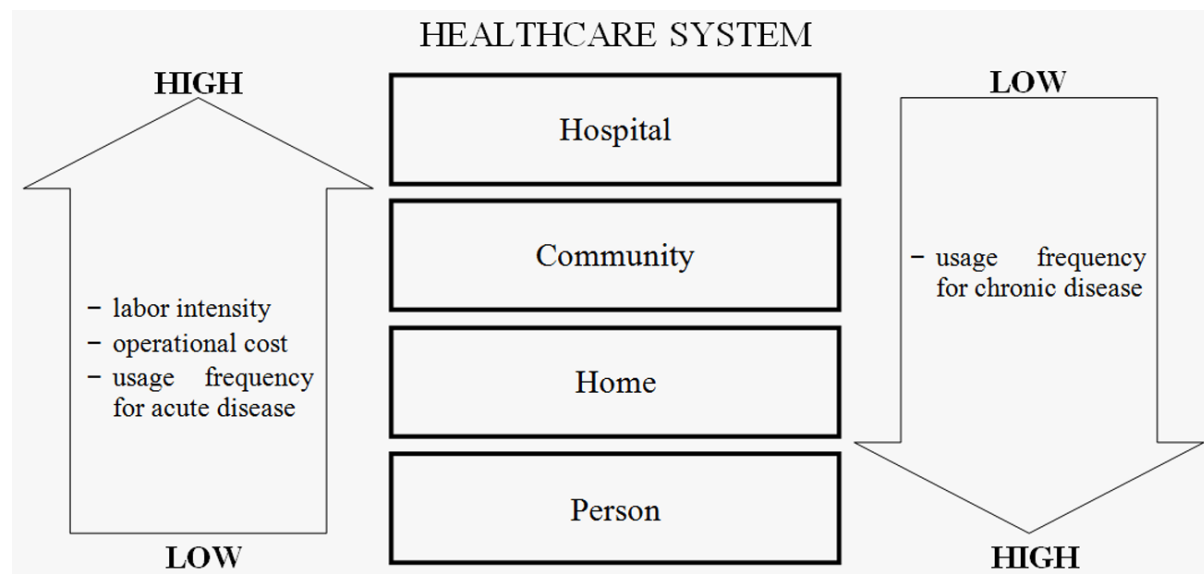
All the aforementioned trends and changes have revealed the necessity to adopt a different model of health care. The modern healthcare provision should be more patient-centric, rather than career-centric, in order to deal effectively with the new challenges. Such a healthcare system should be organized in a layered structure (Pang, 2013) as shown in Figure 1. For example, the lower layer should comprise the single person and the higher one the hospital. Between these two terminal layers one can find the home and community layers. The lowest layer (person) is characterized by the lowest labor intensity and operational cost, as well as the highest frequency of usage for chronic disease and lowest frequency of usage for acute disease. In contrast, the highest layer (hospital) has the highest labor intensity and operational cost, as well as the highest frequency of usage for acute disease and lowest for chronic disease.

The delivery model of health care is now standing at the starting point of its way to transform from the traditional hospital-centric to the modern home-centric/person-centric approach. The Internet of Things with its pervasive, personalized, and ubiquitous character can guide health care through this way. Many relevant concepts have been introduced to describe the future healthcare model powered by emerging information and communication technologies. Various terms, such as connected health, health Internet of Things (Health-IoT), pervasive healthcare (pHealth), ubiquitous healthcare (uHealth), mobile healthcare (mHealth), electrical healthcare (eHealth), telehealth, telemedicine, etc., have been utilized in order to express the plethora of these new concepts (Pang, 2013; Pawar et al., 2012; Rose et al., 2015). All these new concepts include, more or less, the fundamental aspects of the home-centric/person-centric healthcare model.

DOI: 10.4018/978-1-7998-3479-3.ch133

This article presents the concept of the Internet of Things as applied in health care discussing the benefits and opportunities as well as the posed obstacles and challenges. The article focuses on the basic strategies for analyzing healthcare data generated by the Internet of Things enabled devices and describes the current status, the major challenges, and the future trends of this revolutionary field.

Figure 1. The layered structure of a modern healthcare system

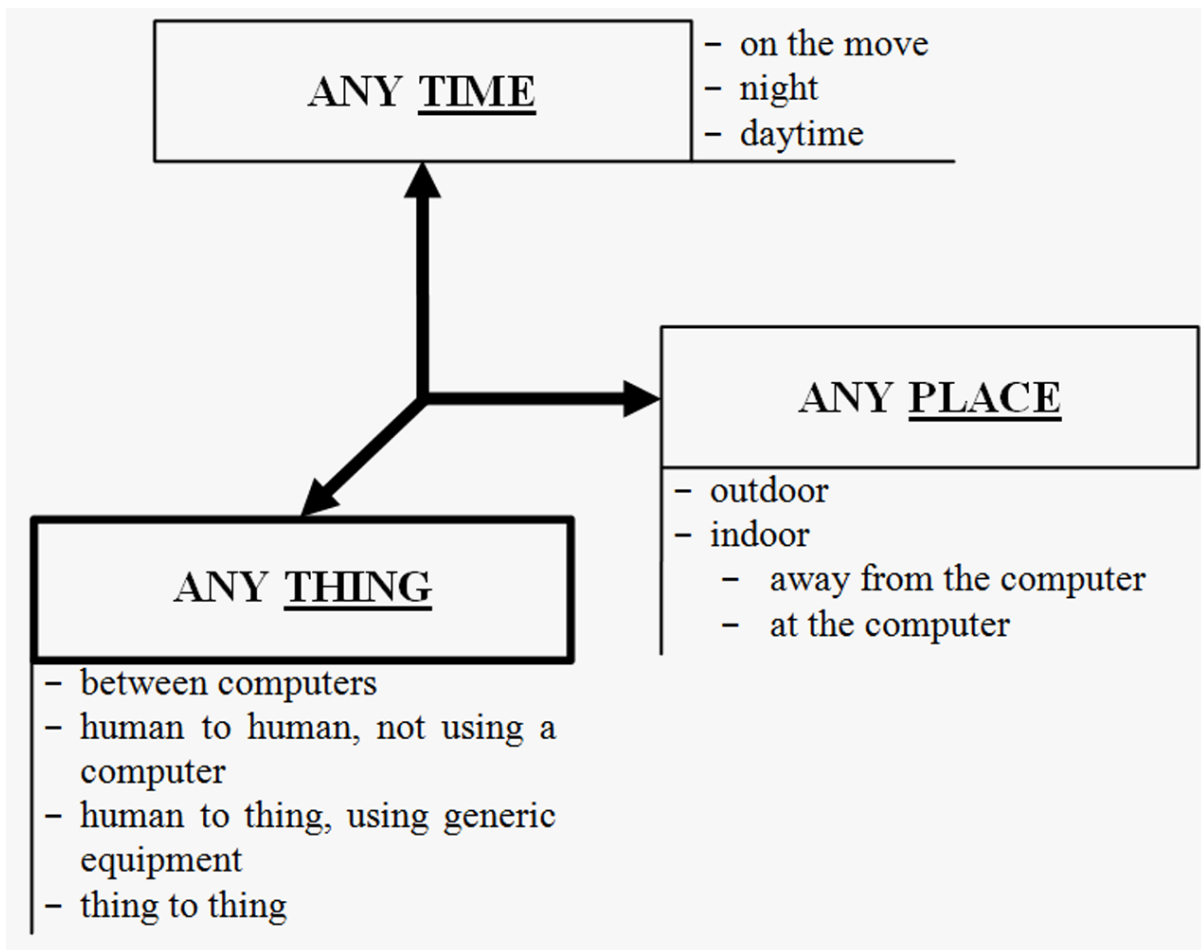


BACKGROUND

According to the definition provided by the telecommunication standardization sector of the International Telecommunication Union (ITU-T Recommendations, 2012), the Internet of Things (IoT) is “a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies”. With regard to the IoT, the term “thing” is defined as “an object of the physical world (physical thing) or the information world (virtual thing), which is capable of being identified and integrated into communication networks”. Exploiting various capabilities, concerning identification, data capture, processing, communication, etc., the IoT uses any “thing” to offer any service to any application domain. At the same time IoT has to abide by the requirements for security and privacy (ITU-T Recommendations, 2012).

The term “Internet of Things” was first used by Ashton in 1999 to refer to the linking of internet and the interconnected objects via radio frequency identification (RFID) technology. (Ashton, 2009). Undoubtedly the IoT is the future of technology, with societal implications as well, that can make everyday life more efficient. The IoT can be recognized as the second digital revolution (The Government Office for Science, 2014), given that the first one was the World Wide Web, which connected people around the world establishing a global community. The IoT is built on the basis of the World Wide Web extending its benefits and opportunities, but poses new big challenges. It is expected to greatly integrate leading technologies, such as technologies related to advanced machine-to-machine communication, autonomic networking, data mining and decision-making, security and privacy protection, as well as cloud computing, with technologies for advanced sensing and actuation. As shown in Figure 2, the IoT

Figure 2. The new dimension (any thing communication) introduced with the Internet of Things (ITU-T Recommendations, 2012)



adds the dimension “any thing communication” to information and communication technologies (ICTs), which already provide “any time” and “any place” communication.

In general, IoT architecture can be described as a four-tiered multilayered structure (Da Xu et al., 2014). From the bottom-up, the first layer is the sensing level, which integrates with all the different types of hardware (RFID, sensors, actuators, etc.) that are connected to the physical world and collect data. The second layer is the networking layer, which provides networking support and data transfer over wired and wireless networks. Next is the service layer which creates and manages all types of services to meet user needs. Finally, the interface layer, which is on top, provides interaction with users and other applications.

Among the plethora of application domains that can benefit by the utilization of the IoT, health care is one of the most important. IoT Analytics, a leading provider of market insights for the Internet of Things (IoT), Machine to Machine (M2M), and Industry 4.0, has identified the top 10 IoT segments in 2018 (IoT Analytics, 2019) and Connected Health is one of them. The other segments include in the list are Smart City, Connected Industry, Connected Building, Connected Car, Smart Energy, Smart Supply Chain, Smart Agriculture, and Smart Retail. Their study utilized data from a structured repository of 1,600 real-world projects that were either ongoing or had been successfully completed. All these projects

were classified as non-consumer IoT implementations. According to IoT Analytics, the most popular IoT segment for 2018 was Smart City, including 367 projects out of 1,600, followed by Connected Industry (265 projects) and Connected Building (193 projects). The geographical distribution of these projects is the following: Americas (45%), Europe (35%), and Asia (16%). Europe includes the majority of Smart City projects (45%), whereas the Americas embrace most Connected Health (55%) and Connected Car projects (54%). Smart Agriculture projects (31%) are most popular in the Asia/Pacific region.

Vilamovska et al. (2009) have classified IoT applications, in particular RFID applications, in healthcare across two axes:

- **IoT enabling functions:** These are four key and mutually exclusive functions concerning assets, staff and patients.
 - Tracking.
 - Identification and authentication.
 - Automatic data collection and transfer.
 - Sensing.
- **Healthcare applications:** These are the four most promising areas of applications in healthcare as identified by Vilamovska et al. (2009).
 - Patient safety and quality of care.
 - Pharmaceutical application (not including supply chain and counterfeit drug issues).
 - Management of devices, supplies and biological material.
 - Patient and healthcare provider support and management.

Examples of healthcare applications for each enabling function are presented as a four-by-four matrix in Table 1.

Table 1. List of the most promising areas for RFID deployment in healthcare (Vilamovska et al., 2009)

	Patient safety/ Quality of care	Pharmaceutical application	Management of devices, supplies and biological material	Patient and healthcare personnel support/ management
Tracking	Tracking vulnerable patients (dementia) Tracking patient location to account for patient time during treatment	Easy recall of products	Prevention of leftins during surgical procedures Equipment tracking (ensure hygiene, compliance, regular maintenance, fast spotting when needed)	Personnel tracking to improve workflow and reduce patient waiting times in emergency room
Identification and authentication	Mother–newborn accurate matching	Auto ID-enabled medication administration system	Maintenance of real-time clinical information associated with patient within a hospital	Patient identification to reduce incidents harmful to patients (wrong drug, dose, time, procedure)
Automatic data collection and transfer	Use of portable devices for care coordination	Tracking pharmaceutical inventories	Inventory management for better use of staff time and faster care delivery	IoT-supported automated care, pathways, procedures audit, and management
Sensing	Patient vital signs to trigger alerts for medical personnel, and remote monitoring at patient's home	Patient compliance with prescribed medication treatment (in and out-patient)	Blood bags equipped with temperature sensors in hospital to ensure cold chain and efficacy	Personnel and asset tracking to ensure infection control (nosocomial infections)

HEALTHCARE DATA ANALYSIS

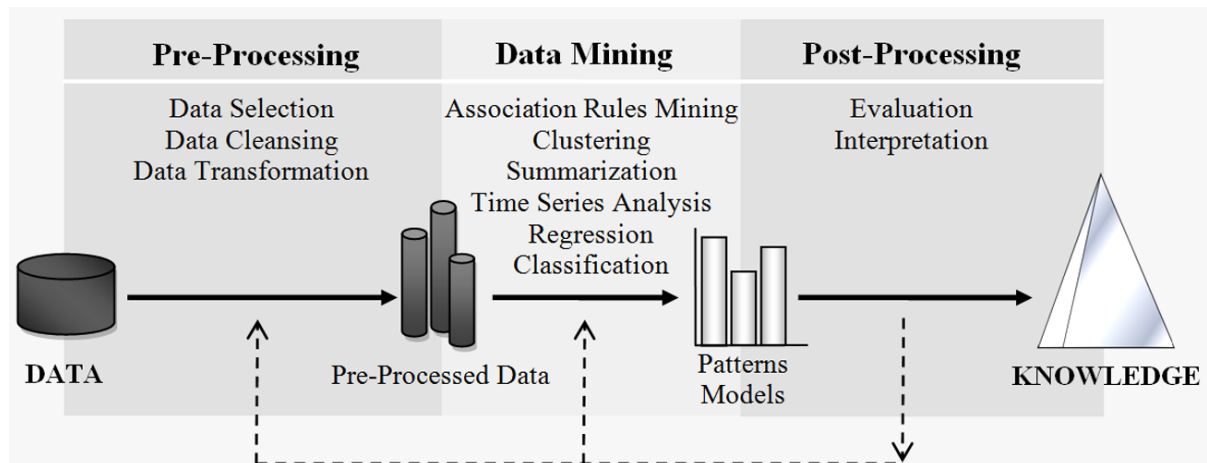
Healthcare data analysis and analytics can support the basic goals of a healthcare system. The objectives that can be addressed through healthcare data analysis have been classified into three major axes by Cortada, et al. (2012) as presented in the following lines:

- Improvement of clinical effectiveness and patient satisfaction
 - Improvement of clinical quality of care
 - Improvement of patient safety and reduction of medical errors
 - Improvement of wellness, prevention and disease management
 - Better understanding of physician profiles and clinical performance
 - Improvement of patient satisfaction, acquisition and retention
- Improvement of operational effectiveness
 - Reduction of costs and increase in efficiency
 - Optimization of network management
 - Improvement of pay for performance and accountability
 - Increase in operating speed and adaptability
- Improvement of financial and administrative performance
 - Increase in revenue and return on investment (ROI)
 - Improvement of utilization
 - Optimization of supply chain and human capital management
 - Improvement of risk management and regulatory compliance
 - Reduction of fraud and abuse

Data Mining

Data mining emerged in order to cope with the challenges that traditional data analysis techniques were facing up when dealing with large amounts of data. Moreover, these data often have a lot of peculiarities (e.g. missing values, noise, etc.). More specifically, data mining is the main step in the process of Knowl-

Figure 3. The KDD Process. Revised from Tzanis (2012)



edge Discovery in Databases or KDD Process (Figure 3). Knowledge Discovery in Databases is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996). However, the term “data mining” is very often used to describe the whole KDD Process. Although the core of the process is the data mining step, where a data mining algorithm is applied in order to extract the patterns from data, the pre-processing and post-processing phases are very important too and contribute sensibly to the quality of the extracted knowledge. The pre-processing phase usually includes the selection of an appropriate portion of data, the cleansing of the selected data, as well as the transformation of data in more appropriate representations. The post-processing phase deals with the management of the produced patterns and models and focuses on the evaluation and interpretation of data mining results.

Data mining, in practice, has the following two “high-level” primary goals (Fayyad et al., 1996):

- **Prediction:** Involves the use of some fields (variables) in a database to predict unknown or future values of other variables of interest.
- **Description:** Focuses on finding human-interpretable patterns describing the data.

Prediction and description are not equivalently important for every data mining application. In the context of Knowledge Discovery in Databases, description tends to be more important than prediction. In contrast, machine learning and pattern recognition applications, usually favor prediction as the primary goal. Prediction and description are achieved by using various data mining tasks, the most common of which are presented in Table 2 (Dunham, 2002). Depending on the nature of the data and the desired knowledge there is a large variety of algorithms for each task. A thorough review of the application of data mining in health care can be found in (Koh & Tan, 2005; Yoo et al., 2012).

Table 2. Most common data mining tasks

Tasks for Description	Tasks for Prediction
<i>Association rules mining:</i> Extracts rules expressing relationships among variables of data. Similar task is <i>sequence analysis</i> where relationships are based on time. <i>Clustering:</i> Groups data, so that data in the same group (cluster) are more similar to each other than to those in other groups. <i>Summarization:</i> maps data into subsets with associated simple descriptions. Also called <i>characterization</i> or <i>generalization</i>.	<i>Classification:</i> Maps data into predefined classes, using a model that is constructed on data with known class mappings. <i>Regression:</i> Maps data into a real valued prediction variable, using a model constructed on data with known mappings. <i>Time series analysis:</i> Examines the sequence of events where an event is determined by one or more of the preceding events.

Big Data Challenges

Although data mining provides the means to deal efficiently with large amounts of, usually, peculiar data, healthcare data analysis in the Internet of Things era is far from an uncomplicated task. The term “big data” is very popular nowadays and is used to describe the explosively extended collections of data, which carry significantly more peculiarities. The popularity of the “big data” concept is signified by the fact that a lot of challenges regarding the management and analysis of big data have been raising, whereas the benefits and opportunities ensuing from the exploitation of these data seem to be manifold and very attractive. The conventional data analysis tools, including machine learning and data mining, are not efficient anymore and have to be scaled in order to deal with the obstacles posed by the big data paradigm.

The basic peculiarities that figure the challenging character of big data are described by the “5-V’s”. Laney (2001) introduced the term “3-V’s” in big data community, namely volume, variety, and velocity. However, since then there have been proposed even more “V’s”, but four of them, variability, value, veracity, and visualization are the most popular (Fan & Bifet, 2012). These “7-V’s” are described in the following lines (Dataconomy, 2014; Fan & Bifet, 2012; Fayyad, 2012):

- **Volume:** The volume of the data has been growing rapidly, more than ever before, providing challenges in loading, processing, and transferring.
- **Velocity:** The velocity of data streams that arrive continuously and need to be analyzed pose challenging real-time constraints.
- **Variety:** There are many different types of data (e.g. text, sensor data, images, audio, video, graph), various degrees of structure in data (structured, semi-structured, unstructured), and often different types and structures of data are mixed.
- **Variability:** The structure of the data as well as the way users are willing to interpret these data usually changes with time providing extra data and knowledge management challenges.
- **Value:** The value refers to the quality of the extracted knowledge that may give the ability of making better decisions and answering more questions.
- **Veracity:** Veracity refers to data accuracy, relating to reliability of data source.
- **Visualization:** Visualization refers to how challenging is data visualization.

SOLUTIONS AND RECOMMENDATIONS

Health care, like many other areas, follows the big data paradigm presented above. Healthcare data are becoming available at rapid rates and as the exploitation of the IoT infrastructure grows, data management and data analysis becomes even more challenging. Subsequently, individual hardware units are not any more adequate to satisfy the computational requirements for managing and analyzing these data. New algorithmic approaches with increased scalability and efficiency and infrastructures that make use of the computational power of multiple hardware units, such as distributed and cloud computing, are deemed necessary.

The prevalent trend in dealing with the big data storm is the concept of cloud computing. According to NIST definition (Mell & Grance 2011), “cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction”.

Cloud services can be grouped in three main groups (Mell & Grance 2011):

- **Software as a Service (SaaS):** The capability provided to the consumer is to use the provider’s applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser or a program interface.
- **Platform as a Service (PaaS):** The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider.
- **Infrastructure as a Service (IaaS):** The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications.

There are various cloud deployment models that have their own benefits and drawbacks. The selection of a proper model depends on the specific application domain. These deployment models are (Mell & Grance 2011; Botta et. al, 2016):

- **Private Cloud:** Provisioned for exclusive use by a single organization and may be owned, managed, and operated by the organization, a third party, or a combination of them.
- **Community Cloud:** Provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns and may be owned, managed, and operated by some of the organizations in the community, a third party, or a combination of them.
- **Public Cloud:** Provisioned for open use by the general public and may be owned, managed, and operated by a business, academic, government organization, or a combination of them.
- **Hybrid Cloud:** Composed of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability.
- **Virtual Private Cloud:** Exploits virtual private network (VPN) technologies for allowing organizations to setup required network settings (e.g. topology, security, etc.).

Cloud computing and the IoT are two novel and promising concepts of technology that have been evolving independently. Although they are different, they have complementary characteristics as shown in Table 3. Due to this complementarity many researchers have been thinking about the possible benefits of the fusion of these two technologies proposing approaches that integrate cloud computing and the IoT (Cloud-IoT) in various application domains (Alhakbaniet al., 2014; Botta et. al, 2016; Diaz et al., 2016; Gomes et al., 2014). The benefits are mutual for these two technologies. The IoT can exploit the virtually unlimited capabilities and resources of cloud computing to compensate its technological constraints and cloud computing can extend its scope to deal with real world “things” in a more distributed and dynamic manner, delivering novel services in a large number of real life cases.

Table 3. Complementary characteristics of cloud computing and IoT (Botta et. al, 2016)

	Internet of Things	Cloud Computing
Displacement	Pervasive	Centralized
Reachability	Limited	Ubiquitous
Components	Real world “things”	Virtual resources
Computational capabilities	Limited	Virtually unlimited
Role of the Internet	Point of convergence	Means for delivering services
Big data	Source	Means to manage

The adoption of the innovative infrastructures of Cloud computing and IoT in health care has lead to a paradigm shift in healthcare data management and analysis. The traditional standalone, local, offline data analysis model has been replaced by a more dynamic, global and regularly online data management and analysis paradigm. Cloud-IoT is able to simplify healthcare processes, enhance the quality of medical services by enabling cooperation among interconnected entities. For example Ambient Assisted Living (AAL) aims at easing the daily lives of people with disabilities and chronic diseases [Botta et. al, 2016].

Cloud-IoT may supply many innovative services, such as patients' vital data collection through sensor networks, data transfer to medical centers' Cloud infrastructures, management, sharing and ubiquitous access of Electronic Healthcare Records (EHR), etc [Botta et. al, 2016, Kuo, 2011]. Moreover, Cloud allows better confrontation with common challenges of healthcare domain such as security, privacy, and reliability, by enhancing medical data security as well as service availability and redundancy. Cloud enables the execution of secure healthcare services, overcoming the issue of limited computational capacity and power. It provides the means for a flexible and scalable storage and processing in order to efficiently perform online and offline analyses of data.

Another new concept in the area of computing is edge computing that can be thought as an advanced version of Cloud computing (Khan et al., 2019). It brings cloud computing services closer to the end-user and is characterized by fast processing as well as short response time. The common practice is that end-users run fast-processing and quick-response applications on their mobile devices that have limited resources, while basic services and data processing are performed on cloud servers. Using cloud services on mobile devices results in high latency and mobility problems. Edge computing meets the aforementioned application requirements by transferring processing to the edge of the network. Thus, cloud computing issues can be solved through the three edge computing models, Cloudlets, Fog computing and Mobile Edge computing (Khan et al., 2019). However, the complexity of Edge computing systems has given rise to a number of challenges, such as mobility management, security and privacy, scalability, heterogeneity, reliability, and resource management (Khan et al., 2019).

FUTURE RESEARCH DIRECTIONS

The introduction of the Cloud-IoT paradigm in the healthcare domain offers numerous benefits and opportunities that will considerably improve the quality of healthcare services in the era of big data. This is a recently emerged research area and is expected to receive the focus of the future research efforts. Although the aims for storage and management of large amounts of data in the IoT infrastructure can be provided by the integration of the IoT and cloud computing, there are still open issues concerning verification, normalization, filtering and analysis of the IoT data (Diaz et al., 2016). The lack of open standards, the large diversity of technologies involved in the IoT, as well as the huge amounts of data require novel techniques to improve and optimize such integration problems.

From a broader perspective, the basic challenges that concern the IoT healthcare applications using cloud and edge computing services and should be addressed in the future are presented below (Atzori et al., 2010; Botta et. al, 2016; Khan et al., 2019; Rose et al., 2015):

- **User's trust, privacy and security:** These challenges may concern the lack of trust in the cloud and edge computing service providers, the lack of knowledge about the physical location of data, the exposure of the distributed system to possible attacks and vulnerabilities, etc.
- **Legal and Social Issues:** The legal challenges are important depending on the specific application and different international laws. Moreover, social issues become extremely important, since humans are becoming the source of sensor data.
- **Large Scale:** This relates to applications which include a very large number of "things", distributed across very wide areas.

- **Reliability and fault tolerance:** This arises with mission-critical applications. Machine learning algorithms for anomaly detection and predictive maintenance are deemed important in order to deal with these problems.
- **Performance:** This concerns applications having specific performance and Quality of Service requirements at multiple levels (communication, computation, storage).
- **Heterogeneity:** This refers to heterogeneity of devices, platforms, operating systems, services, etc.

An emerging trend in the area of IoT is “intelligent connectivity”, which represents the fusion of IoT, Artificial Intelligence and 5G networks promises to provide new transformational capabilities. As defined by GSMA (2018), “Intelligent Connectivity” is the term used to describe “the powerful combination of flexible, high-speed 5G networks, the Internet of Things and Artificial Intelligence. Underpinned by ubiquitous, hyper connectivity, intelligent connectivity marks the beginning of a new era defined by highly contextualised and personalised experiences, delivered as and when one wants them. It will have a significant and positive impact on individuals, industries, society and the economy, transforming how we live and work”.

Given the capabilities of 5G networks, users will be able to wear connected wellness and safety devices that provide ongoing information on their location, vital signs, and other parameters. Emergencies such as falls or catastrophic events can be identified. This information enables individuals to track their specific needs in order to benefit from personalized prevention and predictive healthcare programs, as well as personalized health insurance and safety solutions. As for the healthcare sector, continuous connectivity will enable better management of resources and will ensure that clinics are not run out of equipment and critical medicines. Nowadays, medical expertise is still largely restricted to the location of the physician. However, the advent of tactile Internet will allow the physical examination to be carried out remotely anywhere and at any time, without restrictions (GSMA, 2018).

CONCLUSION

Undoubtedly the IoT is the future of technology, which can provide manifold benefits to health care. However, the posed challenges are also great. Concerning the analysis of healthcare data, various tools have been introduced to deal efficiently with the large volumes as well as the various peculiarities of data (e.g. missing values, noise, etc.). The most popular representative of these modern tools is data mining, or the KDD process, strictly speaking. Although the KDD process has provided a lot of solutions, these techniques have to be scaled in order to deal with the new challenges posed by the big data paradigm. Cloud computing is the modern infrastructure that can provide the means to efficiently manage big data. Both, cloud computing and the IoT are very promising concepts of technology and their complementary characteristics assure that their integration, Cloud-IoT, provides a great potential of applications. The introduction of the Cloud-IoT paradigm in the healthcare domain can offer manifold benefits and opportunities that will considerably improve the quality of health care. The newly emerged area of edge computing complements cloud computing and brings cloud services closer to the end-user and providing faster processing as well as short response times.

REFERENCES

- Alhakhbani, N., Hassan, M. M., Hossain, M. A., & Alnuem, M. (2014). A framework of adaptive interaction support in Cloud-based Internet of Things (IoT) environment, In *Proceedings of the 7th International Conference* (vol. 8729, pp. 136–146). Calabria, Italy: Springer. 10.1007/978-3-319-11692-1_12
- Ashton, K. (2009). Internet of things. *RFID Journal*. Retrieved August 1st, 2019, from <http://www.rfidjournal.com/articles/view?4986>
- Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787–2805. doi:10.1016/j.comnet.2010.05.010
- Botta, A., De Donato, W., Persico, V., & Pescapé, A. (2016). Integration of Cloud computing and Internet of Things: A survey. *Future Generation Computer Systems*, 56, 684–700. doi:10.1016/j.future.2015.09.021
- Cortada, J.W., Gordon, D., & Lenihan, B. (2012). *The value of analytics in healthcare: from insights to outcomes*. IBM Global Business Services, Life Sciences and Healthcare, Executive Report.
- Da Xu, L., He, W., & Li, S. (2014). Internet of Things in Industries: A Survey. *IEEE Transactions on Industrial Informatics*, 10(4), 2233–2243. doi:10.1109/TII.2014.2300753
- Dataconomy. (2014). *Understanding Big Data: The Seven V's*. Retrieved August 1, 2019, from <https://dataconomy.com/2014/05/seven-vs-big-data/>
- Diaz, M., Martin, C., & Rubio, B. (2016). State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing. *Journal of Network and Computer Applications*, 67, 99–117. doi:10.1016/j.jnca.2016.01.010
- Dunham, M. H. (2002). *Data mining: Introductory and advanced topics*. Prentice Hall.
- Fan, W., & Bifet, A. (2012). Mining Big Data: Current Status, and Forecast to the Future. *SIGKDD Explorations*, 14(2), 1–5. doi:10.1145/2481244.2481246
- Fayyad, U. M. (2012). Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. Keynote Talk. BigMine: BigData Mining Workshop KDD-2012, Beijing, China.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press/The MIT Press.
- Gomes, M. M., Da Rosa Righi, R., & Da Costa, C. A. (2014). Future directions for providing better IoT infrastructure. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (pp. 51–54). 10.1145/2638728.2638752
- GSMA. (2018). *Intelligent Connectivity*. Retrieved August 1, 2019, from <https://www.gsma.com/IC/wp-content/uploads/2018/09/21494-MWC-Americas-report.pdf>
- IoT Analytics. (2019). *The Top 10 IoT Segments in 2018 – based on 1,600 real IoT projects*. Retrieved August 1st, 2019, from <https://iot-analytics.com/top-10-iot-segments-2018-real-iot-projects>
- ITU-T Recommendations. (2012). *Overview of the Internet of things*. Recommendation ITU-T Y.2060. Retrieved August 1, 2019, from <http://handle.itu.int/11.1002/1000/11559>

Khan, W. Z., Ahmed, E., Hakak, S., Yaqoob, I., & Ahmed, A. (2019). Edge computing: A survey. *Future Generation Computer Systems*, 97, 219–235. doi:10.1016/j.future.2019.02.050

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2), 64–72. PMID:15869215

Kuo, A. M.-H. (2011). Opportunities and challenges of Cloud computing to improve health care services. *Journal of Medical Internet Research*, 13(3), e67. doi:10.2196/jmir.1867 PMID:21937354

Laney, D. (2001). *3-D Data Management: Controlling Data Volume, Velocity and Variety*. META Group Research Note.

Mell, P. & Grance, T. (2011). *The NIST Definition of Cloud Computing*. Recommendations of the National Institute of Standards and Technology. NIST, U.S. Department of Commerce. Retrieved September 29, 2016, from doi:10.6028/NIST.SP.800-145

Mohammed, E. A., Far, B. H., & Naugler, C. (2014). Applications of the MapReduce programming framework to clinical big data analysis: Current landscape and future trends. *BioData Mining*, 7(1), 22. doi:10.1186/1756-0381-7-22 PMID:25383096

Pang, Z. (2013). *Technologies and Architectures of the Internet-of-Things (IoT) for Health and Well-being* (Unpublished doctoral dissertation). Royal Institute of Technology, Stockholm, Sweden.

Pawar, P., Jones, V., Beijnum, B. F. V., & Hermens, H. (2012). A framework for the comparison of mobile patient monitoring systems. *Journal of Biomedical Informatics*, 45(3), 544–556. doi:10.1016/j.jbi.2012.02.007 PMID:22406009

Rose, K., Eldridge, S., & Chapin, L. (2015). *The Internet of Things: An Overview - Understanding the Issues and Challenges of a More Connected World*. Internet Society White Paper.

The Government Office for Science. (2014). *The Internet of Things: making the most of the Second Digital Revolution*. A report by the UK Government Chief Scientific Adviser. Retrieved September 29, 2016, from <https://www.gov.uk/government/publications/internet-of-things-blackett-review>

Tzanis, G. (2014). Biological and Medical Big Data Mining. *International Journal of Knowledge Discovery in Bioinformatics*, 4(1), 42–56. doi:10.4018/ijkdb.2014010104

Vasanth, K., & Sbert, J. (2012). *Creating solutions for health through technology innovation*. Texas Instruments White Paper. Retrieved September 29, 2016, from <https://www.ti.com/lit/wp/sszy006/sszy006.pdf>

Vilamovska, A.-M., Hatziaandreu, E., Schindler, H. R., Van Oranje-Nassau, C., De Vries, H., & Krapels, J. (2009). *Study on the requirements and options for RFID application in healthcare - Identifying areas for Radio Frequency Identification deployment in healthcare delivery: A review of relevant literature*. RAND Europe.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, 36(4), 2431–2448. doi:10.1007/10916-011-9710-5 PMID:21537851

ADDITIONAL READING

Bui, N., & Zorzi, M. (2011). Health care applications: a solution based on the internet of things. In *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL '11)*. ACM. 10.1145/2093698.2093829

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks*, 11(8), 431047. doi:10.1155/2015/431047

Cloud Standards Customer Council. (2012). *Impact of Cloud Computing on Healthcare*. White Paper. Retrieved August 1, 2019, from <http://www.cloud-council.org/deliverables/CSCC-Impact-of-Cloud-Computing-on-Healthcare.pdf>

Farahani, B., Firouzi, F., Chang, V., Badaroglu, M., Constant, N., & Mankodiya, K. (2018). Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare. *Future Generation Computer Systems*, 78(2), 659–676. doi:10.1016/j.future.2017.04.036

Firouzi, F., Rahmani, A. M., Mankodiya, K., Badaroglu, M., Merrett, G., Wong, P., & Farahani, B. (2017). Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics. *Future Generation Computer Systems*, 78(2), 583–586.

Hou, J.-L., & Yeh, K.-H. (2016). Novel authentication schemes for IoT based healthcare systems. *International Journal of Distributed Sensor Networks*, 11(11), 183659. doi:10.1155/2015/183659

IoT Analytics. (2015). *IoT platforms - The central backbone for the IoT*. White Paper. Retrieved August 1, 2019, from <https://iot-analytics.com/product/iot-platforms-white-paper>

LeSueur, D. (2014). *5 Reasons Healthcare Data Is Unique and Difficult to Measure*. Health Catalyst White Paper. Retrieved August 1, 2019, from <https://www.healthcatalyst.com/5-reasons-healthcare-data-is-difficult-to-measure>

Riazul Islam, S. M., Kwak, D., Kabir, H., Hossain, M., & Kwak, K.-S. (2015). The Internet of Things for Health Care: A Comprehensive Survey. *IEEE Access: Practical Innovations, Open Solutions*, 3, 678–708. doi:10.1109/ACCESS.2015.2437951

Skilton, M., & Hovsepian, F. (2018). *The 4th Industrial Revolution: Responding to the Impact of Artificial Intelligence on Business*, Palgrave Macmillan. Springer International Publishing AG. doi:10.1007/978-3-319-62479-2

KEY TERMS AND DEFINITIONS

Ambient Intelligence: The discipline aims to bring intelligence to everyday environments making them sensitive and responsive to the presence of people in a non-intrusive way and with minimal explicit interaction.

Artificial Intelligence: The area of computer science that studies the construction of computer programs that can imitate human cognitive abilities and present behaviors indistinguishable from human behavior.

Internet of Everything (IoE): A broad term that is used to describe a superset of Internet of Things which includes all sorts of things and connections that one could imagine.

Internet of Medical Things (IoMT): A global infrastructure consisting of the collection of medical devices and applications that are interconnected through information and communication technologies.

Machine Learning: The field of study that deals with the construction of algorithms that can learn from data and make predictions on data.

Radio Frequency Identification (RFID): The technology that incorporates the use of electromagnetic fields to uniquely identify and track a tagged object.

Tactile Internet: The Internet that combines extremely low latency with very high availability, reliability, and security.

Ubiquitous Computing: The discipline aims to bring computing anywhere and anytime in any object. It is also called pervasive computing.

Wi-Fi: The technology that allows electronic devices to connect to a wireless local area network (WLAN) and Internet using radio waves.

Wireless Sensor Network (WSN): The network of spatially distributed autonomous sensors which monitor physical or environmental conditions and cooperatively pass their data through the network to a main location.

Internet of Things in Healthcare as an Innovative Form of Personalized Medicine

16

Ljubica Dikovic <https://orcid.org/0000-0002-4528-4014>*College of Applied Sciences Uzice, Serbia*

INTRODUCTION

According to the US National Intelligence Council, there are six technologies with potential impacts on the US interests out to 2025 (the US National Intelligence Council, 2009): Biogerontechnology; Energy Storage Materials; Biofuels and Bio-based Chemicals; Clean Coal Technologies; Service Robotics; Internet of things.

The great potential offered by the Internet of Things technology enables their wide applications in many areas of society, which would significantly increase and improve the quality of their functioning. By equipping various environments, i.e. domains, even with devices with primitive intelligence and modest communication capabilities, the communication of these entities with each other would be possible, with an aim to ensure data management. Such systems can be widely used in the following areas: Healthcare Domain; Smart Environment Domain; Personal and Social Domain; Transport and Logistics.

BACKGROUND

The Internet of Things (IoT) refers to wireless networks between objects (things). ‘Things’, i.e. objects, become entities with virtual properties which operate and communicate in smart spaces using intelligent interfaces.

Also, the “Internet of Things” is the general idea of things, especially everyday objects that are readable, recognizable, locatable, addressable, and controllable via the Internet - either via Radio Frequency Identification (RFID), Bluetooth, Wi-Fi, telephonic data services, wide-area network, or other means (the US National Intelligence Council, 2009).

In their research paper, Atzori et al. (2010) state that the Internet of Things can be realized in three paradigms: internet-oriented (middleware), things-oriented (sensors) and semantic-oriented (knowledge).

Over the last 20 years, continuous changes in the healthcare domain have taken place, caused by the wide use of information and communication technologies in the medical field. IoT plays a significant role in the broad range of healthcare applications which could be grouped as follows (Atzori et al. (2010)):

- Tracking of Objects and People (Staff and Patients);
- Identification and Authentication of People;
- Automatic Data Collection and Sensing.

DOI: 10.4018/978-1-7998-3479-3.ch134

The rapid growth of IoT has resulted in a massive growth of data generated by these devices and sensors put on the Internet. The physical-cyber-social big data consist of these IoT data, complemented by the relevant Web-based and social data (Sheth, 2016).

The Internet of Things has been identified as one of the emerging technologies in the IT field. The market adoption of IoT has been forecast to take 5–10 years (Gubbi et al. (2013)).

The Internet of Things (IoT) is a platform and a new paradigm for interconnecting electronic devices i.e. global communication among millions of electronic devices connected to the internet.

Personalized healthcare systems deliver e-health services to fulfill the medical and assistive needs of the aging population. Internet of Things is a significant advancement in the Big Data era, which supports many real-time engineering applications through enhanced services. (Jagadeeswari et al. (2018), Khan et al. (2018)).

E-health in Serbia has been the subject of some studies. It was concluded that information and communication technologies are rarely implemented (Milenkovic et al. (2012)).

IoT is a vision which refers to the humanization of technology.

IoT IN HEALTHCARE DOMAIN

Medical sensors are devices that measure a number of physical, chemical, or biological parameters and then transmit or report these data. Some sensors are designed to work outside the body, and others are implanted in the body.

In the healthcare domain, the Internet of Things in its essence covers the following (Li et al. (2009)):

- Body Area Network (BAN): the thermometers, smart t-shirts, smart devices and sensors for health, paper-based home pregnancy tests, etc. supporting personal medical treatment and healthcare.
- Wireless Body Area Network (WBAN): supporting remote medical treatment and healthcare.
- Local Area Network (LAN): the wireless access-based remote patient monitoring system, the smart devices as a hospital interface, pulse oximeter also known as the blood-oxygen monitor, etc.
- Wide Area Network (WAN): telemedicine solutions, distance medicine, etc.
- Very Wide Area Network (VWAN): the smart healthcare solutions as e-health services everywhere, no longer tied to physical locations.

Tracking involves memorizing the positions in real time (e.g. patient flow monitoring) in order to improve work processes in hospitals, but can also refer to the general tracking of motion (e.g. the access to certain rooms, warehouses or inventory material tracking such as blood samples and other materials necessary for the hospital functioning). Identification of patients includes activities, actions and tasks aimed at reducing incidents harmful to patients (e.g. wrong drug, dose, time, and procedures), the maintenance and updating of the complete medical archives in the digital form. Identification and authentication include protocols and procedures regarding the work of medical staff, which promotes morals and ethics of employees in order to increase patient safety. The automatic data collection and transfer involves the automation of various processes as well as the management of medical supplies. This function is supported by the integration of the Radio Frequency Identification (RFID) technology and other health information and technologies. RFID is a method of storing and downloading data remotely, using devices carrying tags. From the RFID group's point of view, the Internet of Things is the worldwide network of interconnected objects uniquely addressable based on the standard communication protocols

(Gubbi et al. (2013)). Sensor devices allow more successful diagnosis of patients in real time; provide various solutions in the field of telemedicine, warnings about the patient's condition and so on. They offer numerous opportunities that would provide, for example, a wireless access-based remote patient monitoring system, which would be designed so as to accompany the patient everywhere in real time, thus enabling continuous monitoring of the patient's bio-signals (Niyato et al. (2009)).

CONCEPT OF M-HEALTH

As mobile technology becomes more affordable and widespread all over the world, using the mobile-enabled technology when providing healthcare services has great potential to improve the health of millions.

Mobile Health, mHealth is a term used for medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, etc.

MHealth is a relatively new concept, but its main advantage is that it enables healthcare professionals to collect data about their patients at any time. The data is processed, analyzed and wirelessly communicated to a gateway device over a cloud. (Paschou et al. (2013), Paschou et al. (2015), Istepanian et al. (2004)).

Considering that MHealth, as a part of eHealth, can be defined as the synergistic solution of mobile computing, a medical sensor, and communications healthcare technologies, whose common goal is wireless monitoring of the psycho-physical state of health or remote patient monitoring. In the Republic of Serbia, electronic health cards are a pioneering venture aimed at promoting mobile healthcare services (Milenkovic et al. (2012)).

An MHealth solution usually consists of a unit designed for patients, which is composed of sensors for medical measurements and accessories, as well as a communication device for 2G and 3G communication, which forwards the data to the health system via cloud. Because of their miniature dimensions, high performance and low cost/price, different sensors, which can be easily wirelessly connected, are used for the measurement of chemical, biological and physical parameters in the body. Also, there are a lot of web-based platforms that provide additional opportunities for the expansion of social networks for different purposes (e.g. health education of patients worldwide). Even more, the network of sensors can be connected via patient's mobile phone to a server that is located in a medical institution. The data collected by the sensor network can be periodically or in real time sent to the server, which approaches the patient's doctor suggesting corrective actions via e-mail or SMS. The program can be set so as to, if the sensor encounters some irregularities, immediately notify a doctor, who could react momentarily.

Thanks to this approach, patients with chronic diseases can be under continuous remote supervision at home. In addition to the increased safety of the patient, it represents significant savings of hospital resources, and hospital beds remain available for other patients who need them. Also, this solution helps hospitals manage beds efficiently and consequently decreases hospital costs. Hospitalized patients whose health status requires full attention can be constantly monitored using IoT-driven monitoring (sensors).

The market division is as follows:

- Preventive care tools which help people stay healthy (e.g. Behavioral change, Community communication, Education).
- Curative care tools as new efficient ways to treat patients (e.g. Remote monitoring, Assisted living, Tele-consultations).
- Productivity tools which improve efficiency and availability (e.g. Electronic health records, Professional communication).

TELEHEALTH

Telehealth (telemedicine) is a health delivery system which refers to the exchange of medical information between two sites via electronic communications in order to improve a patient's health status. Telehealth includes a growing range of applications and services using two-way video telephony, wireless tools and networks, emails, smart devices, Internet and cloud services and infrastructures, and other forms of telecommunications technology. Clinical telemedicine practice supported by video conferencing, medical connectivity through m-health, remote patient monitoring, disease management, home healthcare, continuing medical education, medical call centers, are all considered part of telemedicine and telehealth. Health professionals providing telehealth services should have the continuing education, training, licensure, in order to acquire the necessary knowledge and competencies needed to provide safe and high-quality telehealth services.

From 2013 to 2020 the telehealth market's compound annual growth rate is expected to be around 46 percent. The digital health market is expected to reach 233.3 billion US dollars by 2020. The telemedicine market has grown significantly in recent years. As of 2015 the telemedicine global market was valued at some 18 billion U.S. dollars. The market is expected to grow significantly by 2021 and is expected to be valued at more than 41 billion U.S. dollars at that time. The statistic shows the projected growth of the total number of telehealth patients worldwide from 2013 to 2018. Also, a statistic show the estimated number of mHealth app downloads worldwide from 2013 to 2017, in billions of downloads. It is estimated that in 2017 there will be 3.7 billion mobile health app downloads (Statistic (n.d.)).

How telepsychiatry expands access to care? In February 2018, American Psychiatric Association (APA) updated its Policy on Telepsychiatry: "Telemedicine in psychiatry, using video conferencing, is a validated and effective practice of medicine that increases access to care. The American Psychiatric Association supports the use of telemedicine as a legitimate component of a mental health delivery system to the extent that its use is for the benefit of the patient, protects patient autonomy, confidentiality, and privacy; and when used consistent with American Psychiatric Association policies on medical ethics and applicable governing law." (Telepsychiatry. (n.d.))

With the "anytime, anywhere framework," telepsychiatry is changing the way communities and the medicine as a whole approach mental and behavioral healthcare.

Telepsychiatry delivered via videoconferencing technology allows patients to connect with providers remotely from their home or other private place where they are located.

By increasing access to care and real time support with qualified professionals who can quickly and effectively assess situations, with a proactive intervention that minimizes the potential for escalation, telepsychiatry is becoming an important strategy in the future in preventing the mental health crisis.

Telepsychiatry helps to avoid unnecessarily transferring patients to a hospital by instantly connecting to behavioral health specialists to assess and determine the best treatment plan.

Telepsychiatry plays an important role as a solution that eliminating unnecessary hospitalization, opens up a new approach to mental illness, and empowers timely and more effective mental and behavioral healthcare.

TECHNICAL REQUIREMENTS FOR REAL-TIME INTERACTIVE SESSIONS

The quality of a video healthcare interactive session primarily depends on the characteristics of the network connection between the conferencing sites. The H.323 standard specifies the components, pro-

protocols, and procedures for the transmission of real-time audio, video, and data communications over IP (Internet Protocol), i.e. packet-based, networks. In the case of H.323, a high-quality conference, which implies excellent audio and video, needs from 768 kbps (kilobits/second) to 2 Mbps (megabits/second) in each of the downlinks and uplink directions.

In the case of “Point to point video conferencing”, two H.323 video conferencing systems communicate to each other over the Internet. One station initiates the call, and the other either accepts or rejects it. Once accepted, encoded and compressed audio and video flow between the two stations. In the case of “Multipoint video conferencing” more than two sites communicate with each other. In the H.320/ISDN world, a Multipoint Control Unit (MCU) was used. This allowed three or more user stations to connect to the MCU. In the case of the voice switching mode, the MCU made a decision about which site was currently talking. Smartphones and tablets (such as iPhones and iPads) do support live conferences; however, 3G/4G/LTE/WiFi connections are not sufficiently reliable for participating in online sessions. It is highly recommended to use a computer with a wired internet connection. In the case of high-quality (HQ) resolution, 640x360@15, 30 fps (frame per second) is recommended, but in the case of full High Definition resolution (FHD), 1920x1080 30 fps is recommended. Graphical cards must be compatible with DirectX 10.0, from 512MB (min. NVidia 8600 GS, Radeon HD 2600 XT, Intel HD Graphics 2000) to 2GB (min. NVidia GT640, AMD Radeon HD 6670, Intel HD Graphics 4000).

Minimum system requirements for conference hosts are: Windows XP 32 bit or Windows 7 (32bit or 64bit); Core2 Quad, Intel i3, i5 or i7 equivalent CPU; 4 GB of RAM; 100 MB of available HDD space (additional space recommended for conference recording features); 100 MB NIC or higher; High speed broadband Internet access (at least 1 Mbps upstream and 4 Mbps downstream); DirectX 9.1 or later compatible webcam; DirectSound compatible audio input and headset or speakers (Video Conferencing System Requirements (n.d.), Web & Multiparty Video Conferencing System Requirements (n.d.), VCI - Next Generation Video Conferencing (n.d.)).

The video conference software should be able to adapt to changing bandwidth environments without losing the connection. Health organizations should have the appropriate redundant systems in a place that ensures availability of the data transmission infrastructure for critical connectivity.

DIGITAL TRANSFORMATIONS IN MEDICINE

mHealth has the potential for use in the following areas:

- Education and training of healthcare workers;
- Remote data collection and remote monitoring;
- Remote diagnostic and treatment support.

Today's healthcare consumers are more motivated than ever before to take action for better health. In response, healthcare organizations from hospitals to large health systems, are seeking methods to ensure new tactics to meet the core strategic need of the healthcare industry: a higher value of care at a lower cost. As such, three key trends are developing as consumerism takes hold in healthcare. Key learning objectives: (i) Leveraging data from consumer health technologies to better engage patients; (ii) Creating a provider-patient dialogue utilizing patient-generated health data; (iii) The value of deploying enterprise platform solutions for care management (Validic. (n.d.)).

The main advantages brought by the concept of mHealth imply the increase in the quality of health care, while creating new opportunities on the one hand, and reduced costs on the other hand. About 75% of patients suffer from chronic diseases (high blood pressure, diabetes, heart disease, cancer, etc.). Using mHealth solutions, the biomarkers of these chronic patients can be monitored ensuring the quality management of their diseases in order to improve the quality of their life. These technologies exist and have been adopted, and their examples are: the wireless Helder monitors, weight scales, blood pressure cuffs and glucometers (Malvey et al. (2014)).

In order to be cost-effective and commercially-sustainable for the mass-market, mobile health services should be standardized. One of the problems is that today there is not an organization in the world which deals with the standards in the field of mHealth. For example, the International Standards Organization (ISO) has developed standards (IEEE 11073) for the transmission of blood pressure in a basic binary data format between two low-level devices, which can be transferred into a human-readable format using health messaging standards, such as HL7 (Engineering in Medicine and Biology Society (n.d.)).

Some of the mHealth categories (Applications and Devices) are: Nutrition Application, Glucometers, Biometrics devices, Sleep Trackers Applications and devices, Tobacco Cessation Applications and Programs, Medication Adherence Devices and so on (Validic. (n.d.)).

Some examples of the use of medical sensors are as follows: the sensors can help detect, and thus avoid, about 5000.00 known viruses that cause serious illness around the world. Biochemical sensors can continuously scan a billion miles long chain of Deoxyribonucleic acid (DNA) and look for changes and potential diseases. Taking into account the fact that the human heart throughout a life pumps about 15 million gallons of blood, cardiac sensors or heart rhythm sensors can provide information about how efficiently a heart works. The human brain has about 100 billion neurons. Gum sensors have the preventive effect on the development of other diseases in the organism because people who suffer from a gum disease represent 50% of those who are predisposed to suffer brain and heart attacks. New sensors can detect a high level of blood sugar without drawing blood. By using an ultra-sensitive sensor to detect the presence of tumors and early abnormal cell growth, cancer can be recognized at an early stage and thus effectively treated. Biosensors may provide insights into disease processes that are hard to detect directly, such as dysfunctions in brain chemistry that are thought to play an important role in many mental disorders (Health Sensing Infographic, (2014), Sensors. (n.d.)).

The next generation of IoT quantified tracking devices falls into the category of wearable electronics and multi-sensor platforms. These products include smart watches, wristband sensors, wearable sensor patches, artificial reality-augmented glasses, brain-computer interfaces, wearable body metric textiles.

Other important categories include smart phone applications and their enhancements, and environmental monitoring and home automation sensors.

For example, Hexoskin products give insight into the physical training, sleep, and personal daily activities:

- Track heart rate and view real-time ECG with great precision.
- Heart rate variability (HRV) is an essential tool to measure stress and training fatigue to avoid overtraining.
- Breathing rate in order to control own breathing, increase performance and reduce stress.
- Breathing volume in order to control lung capacity for each of daily activities and avoid hyperventilation and apnea.
- Sleep tracking device in order to measure resting heart rate, breathing, and sleep positions.

How could smart pills be innovating health care?

Health care providers are faced with inconsistent and irregular intake of drugs, especially in patients with chronic diseases. With the new phone applications, alarms, smart packaging and IoT innovations such as “smart pill”, the problem of regular patient medication can be improved.

The wireless health technology is beginning with smart pill technologies, also known as “ingestible”, which could change the way of taking the therapy. Key benefits include maximizing drug effectiveness and reducing medical costs due to improper drug usage.

A pill is considered “smart” if it has technological components such as microchips, cameras or sensors that communicate wirelessly with wearable software or mobile applications, on dislocated computers in medical offices, as well as with smart devices.

A smart pill, also known as digital pill, has a drug and an ingestible sensor. The sensor is activated when it comes into contact with stomach fluid to detect when the pill has been taken. The data is then transmitted to a wearable sensor patch which may possibly be transmitted the information to a paired smartphone application. Health care providers, with the patient’s consent, can then access the data via a web portal.

Further expansion of support for digital medicines into the oncology treatment area is not only important for patients and service providers; it will be going in the direction of change in the pharmaceutical industry developing therapies intended to one day eradicate cancer. For example, data gathered from digital oral oncolytic will enable cancer drugs and treatment regimens to be optimized to work their best for each individual patient, something not possible until now. The mechanism which turns data into predictive knowledge as a part of artificial intelligence could help healthcare professionals better understand a person’s health and could have a significant impact on improving quality of life and well-being (Proteus Trends (n.d.)).

In a world where everything is becoming connected, the development of this small piece of technology that can communicate in the field of human health and share information about a patient’s condition with health care providers has received a lot of scientific attention.

It should be kept in mind that these new technologies bring new dangers. Such devices, like all computers, can be hacked or can be broken. Questions about who should control these devices in the body and how they might be turned off are also very important. The advent of these digital pills brings many of ethical quandaries. These drugs and devices hold implications for transforming the clinician-patient relationship, patient trust, patient monitoring, clinical management, and risk management. As these medical technologies become more widespread, manufacturers and providers of health services will need to develop mechanisms to maximize patient benefits and minimize the risks.

FUTURE RESEARCH DIRECTIONS

Generally, the next Internet revolution will be the interconnection between everyday existing objects in order to create a smart grid and intelligent environment. The future application of technology in health care will lead to the creation of an entirely new level of personalized, digital health care, where everyone is responsible for monitoring their own health and the quality of their own life. Further research is aimed at improving the existing sensors through increasing their capabilities and enhancing their efficiency. In the coming period, IoT is expected to play the key roles in all the aspects of modern medical treatment and health management - prevention, diagnosis, disease monitoring, treatment monitoring. This paper emphasizes the growing needs for better functioning of healthcare systems in real-time as well as the

future development of personalized medicine. This innovative technology in the future offers new opportunities for patients, including improvements in lifestyle of patients, and new opportunities for health systems, including reduction of healthcare costs.

From the academic perspective, technological research in data science, machine learning and artificial intelligence will continue to grow. At the social level, we have to consider many ethical issues concerning privacy policy, authority and autonomy that intelligent devices will have.

CONCLUSION

The advances in science, technology, engineering, and medicine will enable the use of sophisticated sensor technologies in the further medical research that could dramatically improve the quality of life of people with serious disabilities as well as in the field of preventive medicine. We can conclude that the technological advancement in the healthcare area revolves around the following outcomes: improved access to care, increased care quality and reduced care costs. The business models behind IoT have been established around the principles of cloud-assisted IoT, where devices collect data that is then transmitted to cloud services. There is, of course, the major issue of ethical use and distribution of sensitive data, which applies to the whole digital world but is especially important in IoT healthcare systems. The Internet of Things is potentially disruptive because global society, global economy and governments are unprepared for a possible future when the Internet nodes will be able to remotely control, locate, and monitor everyday things. Having in mind that everyday objects become information security risks, the IoT could distribute those risks far more widely than the Internet has to date.

Also, based on today's technology, there is simply not enough wireless spectrums to connect trillions of IoT. At this point, the issue of generating and distributing a sufficient amount of energy for the operation of all IOT devices is unresolved, as well as the problem of storing and processing huge amounts of IoT data. To solve these issues, new scientific discoveries are required, especially in low-power communications, energy saving, process optimizations and distributed machine learning.

REFERENCES

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787–2805. doi:10.1016/j.comnet.2010.05.010

Biometric shirts for performance improvement and sleep tracking. (n.d.). Retrieved March 28, 2016, from <http://www.hexoskin.com/>

Clim, A., Zota, R. D., & Tinica, G. (2018). Big Data in home healthcare: A new frontier in personalized medicine. Medical emergency services and prediction of hypertension risks. *International Journal of Healthcare Management*. Advance online publication. doi:10.1080/20479700.2018.1548158

Global digital health CAGR by segment 2013-2020 | Statistic. (n.d.). Retrieved March 7, 2019, from <http://www.statista.com/statistics/387875/forecast-cagr-of-worldwide-digital-health-market-by-segment/>

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645–1660. doi:10.1016/j.future.2013.01.010

Health Sensing Infographic. (2014). Retrieved March 28, 2016, from <http://sensing.xprize.org/about/health-sensing-infographic>

Istepanian, R., Jovanov, E., & Zhang, Y. (2004). Guest Editorial Introduction to the Special Section on M-Health: Beyond Seamless Mobility and Global Wireless Health-Care Connectivity. *IEEE Transactions on Information Technology in Biomedicine IEEE Trans. Inform. Technol. Biomed.*, 8(4), 405–414. doi:10.1109/TITB.2004.840019

Jagadeeswari, V., Subramaniaswamy, V., Logesh, R., & Vijayakumar, V. (2018). A study on medical Internet of Things and Big Data in personalized healthcare system. *Health Inf Sci Syst*, 6(1), 14. doi:10.1007/13755-018-0049-x PMID:30279984

Khan, M., Han, K., & Karthik, S. (2018). Designing Smart Control Systems Based on Internet of Things and Big Data Analytics. *Wireless Personal Communications*, 99(4), 1683–1697. doi:10.1007/11277-018-5336-y

Khan, M., Silva, B. N., & Han, K. (2017). A Web of Things-Based Emerging Sensor Network Architecture for Smart Control Systems. *Sensors (Basel)*, 17(2), 332. doi:10.3390/17020332 PMID:28208787

Li, H., Takahashi, T., Toyoda, M., Mori, Y., & Kohno, R. (2009). Wireless Body Area Network Combined with Satellite Communication for Remote Medical and Healthcare Applications. *Wireless Pers Commun Wireless Personal Communications*, 51(4), 697–709. doi:10.1007/11277-009-9765-5

Malvey, D., & Slovensky, D. J. (2014). From Telemedicine to Telehealth to eHealth: Where Does mHealth Fit? *mHealth*, 19–43.

MHealth Ecosystem | Validic. (n.d.). Retrieved March 5, 2019, from <http://validic.com/developers/mhealth-ecosystem>

Milenkovic, D., Jovanovic-Milenkovic, M., Vujin, V., Aleksic, A., & Radojicic, Z. (2012). Electronic health system: Development and implementation into the health system of the Republic of Serbia. *Military Medical and Pharmaceutical Journal of Serbia VSP*, 69(10), 880–890. PMID:23155610

Niyato, D., Hossain, E., & Camorlinga, S. (2009). Remote patient monitoring service using heterogeneous wireless access networks: Architecture and optimization. *IEEE J. Select. Areas Commun. IEEE Journal on Selected Areas in Communications*, 27(4), 412–423. doi:10.1109/JSAC.2009.090506

Online N. Web & Multiparty Video Conferencing System Requirements. (n.d.). Retrieved March 27, 2016, from <http://www.nefsis.com/Support-Video-Software/conferencing-system-requirements.html>

Paschou, M., Papadimitiriou, C., Nodarakis, N., Korezelidis, K., Sakkopoulos, E., & Tsakalidis, A. (2015). Enhanced healthcare personnel rostering solution using mobile technologies. *Journal of Systems and Software*, 100, 44–53. doi:10.1016/j.jss.2014.10.015

Paschou, M., Sakkopoulos, E., Sourla, E., & Tsakalidis, A. (2013). Health Internet of Things: Metrics and methods for efficient data transfer. *Simulation Modelling Practice and Theory*, 34, 186–199. doi:10.1016/j.simpat.2012.08.002

Sensors. (n.d.). Retrieved March 28, 2016, from <https://www.nibib.nih.gov/science-education/science-topics/sensors>

Sheth, A. (2016). Internet of Things to Smart IoT Through Semantic, Cognitive, and Perceptual Computing. *IEEE Intell. Syst. IEEE Intelligent Systems*, 31(2), 108–112. doi:10.1109/MIS.2016.34

Slovensky, D. J., & Malvey, D. M. (2017). Introduction to Focused Issue on mHealth Infrastructure: issues and solutions that challenge optimal deployment of mHealth products and services. *mHealth*, 3(52). <http://mhealth.amegroups.com/post/view/mhealth-infrastructure-issues-and-solutions>

Trends, G. (n.d.). Retrieved May 7, 2019, from <https://www.google.com/trends/>

VCI - Next Generation Video Conferencing. Blog. (n.d.). Retrieved March 27, 2016, from <http://www.vci.no/portal/Blog.aspx>

Video Conferencing System Requirements. (n.d.). Retrieved March 27, 2016, from <http://trueconf.com/support/system-requirements.html>

Your Global Connection to the Biomedical Engineering Community - Engineering in Medicine and Biology Society. (n.d.). Retrieved Jun 17, 2019, from <http://www.embs.org/>

ADDITIONAL READING

Appari, A., & Johnson, M. E. (2010). Information security and privacy in healthcare: Current state of research. *International Journal of Internet and Enterprise Management IJIE*M, 6(4), 279. doi:10.1504/IJIE.2010.035624

Babar, M., & Arif, F. (2018). *J Ambient. Intell Human Comput.*, doi:10.1007/12652-018-0820-5

IoT Reference Design Kit. (2015). *For Future Smart Connected Life and Business Collaborative Internet of Things (C-IOT)*, 209-223.

Kim, J. T. (2014). Privacy and Security Issues for Healthcare System with Embedded RFID System on Internet of Things.

Maksimovic, M., Vujovic, V., & Perisic, B. (2015). A custom Internet of Things healthcare system. *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*.

Mukherjee, S., Dolui, K., & Datta, S. K. (2014). Patient health management system using e-health monitoring architecture. *2014 IEEE International Advance Computing Conference (IACC)*. 10.1109/IAdCC.2014.6779357

Po, S. N., Dagang, G., Hapipi, M. D., & Hock, F. T. (2006). MEMSWear - Biomonitoring - Incorporating sensors into smart shirt for wireless sentinel medical detection and alarm. *J. Phys.: Conf. Ser. Journal of Physics: Conference Series*, 34, 1068–1072. doi:10.1088/1742-6596/34/1/176

Telepsychiatry. (n.d.). Retrieved Jun 17, 2019, from <https://www.psychiatry.org/psychiatrists/practice/telepsychiatry>

Trends, P. (n.d.). Retrieved Jun 17, 2019, from <https://www.proteus.com/discover/>

Villar, R., Beltrame, T., & Hughson, R. L. (2015). Validation of the Hexoskin wearable vest during lying, sitting, standing, and walking activities. *Appl. Physiol. Nutr. Metab. Applied Physiology, Nutrition, and Metabolism*, 40(10), 1019–1024. doi:10.1139/apnm-2015-0140 PMID:26360814

Xiao, Z. R., Lv, B. G., Wang, X., & Zhao, Y. J. (2013). A Healthcare Service System Based on Internet of Things. *AMR Advanced Materials Research*, 774-776, 1903–1907. doi:10.4028/www.scientific.net/AMR.774-776.1903

KEY TERMS AND DEFINITIONS

Body-Area Network: A network that can collect information about an individual's health, fitness, and energy expenditure.

Health Informatics (Healthcare Information Systems): Software solutions for appointment scheduling, patients' data management, work schedule management and other administrative tasks regarding health.

Healthcare Monitoring: The medical applications can be of two types: wearable and implanted. Wearable devices are used on the body surface of a human or just at a close proximity of the user. The implantable medical devices are those that are inserted into the human body.

Internet of Things (IoT): A new concept that has emerged thanks to the development of new technologies, especially the Internet Technologies and Wireless Sensor Networks that together enable the existence of the health system at the global level.

Mobile Health (mHealth): A collection of mobile technologies as tools and platforms for health research and healthcare provision.

Sensor Network: A network that consists of multiple detection stations called sensor nodes, each of which is small, lightweight, and portable.

Sensor Node: A node in a sensor network that is capable of performing some processing, gathering of sensory information and communicating with other nodes connected in the network. The main components of a sensor node are a microcontroller, transceiver, external memory, power source and one or more sensors.

Telemedicine: A physical and psychological diagnosis and treatment at a distance, including tele-monitoring of patients' functions.

Wireless Sensor Networks (WSN): A network that consists of spatially distributed autonomous sensors used to monitor physical or environmental conditions, and to forward their data jointly to the main location through the network.

Highly Aged and After Cloud

Shigeki Sugiyama

 <https://orcid.org/0000-0003-3764-4160>

Independent Researcher, Japan

INTRODUCTION

There have been living more and more highly aged people (after 75 years old). And the number of those highly aged people has been increasing year by year in Japan (JILI, 2019). This might be also a global phenomenon. At this age, some of the tissues and the organs of human will seem to start functioning not perfectly well (Yomeishu, 2012. Tentouyobounavi, 2017). And they will seem to be fragile by all kinds of inside and outside stresses. In another words, it seems that some kinds of phase state changes of the human body will make themselves physically weak. However, it is also known the following things too;

1. A bit of mental assistance will make them work fairly good again.
2. A little bit of physical help will be possibly to make them work very good again. So that, one will not need another's help (including a doctor's help) any more.
3. Just a bit of rest will make the physical body and the mental damage work well again.
4. A novice medical treatment will make the physical body work well again.
5. Just a bit of massage and tapping will make the physical body and the mental damage work again.
6. Etc.

So, in this chapter, it will be studied what sort of things can be done for highly aged people by the help of present high technology of "Cloud". What is more, it will be considered whether "Cloud" is only the technology for helping and assisting the highly aged people or not? Namely, human will understand a thing better if there is an assistance visually and intimately. In this sense, the following technologies are also considered for the usages; Individual Cloud, Fog Computing, AI (Artificial Intelligence), VR (Virtual Reality), and AR (Augmented Reality).

The studies and the arguments will go on under the following steps.

1. Firstly, it will be considered what sort of physical body and mental damages will come out to the highly aged people.
2. Secondly, it will be studied what sort of assistance and help will be needed for those damages.
3. Thirdly, it will be studied what sort of assistance and help will be possible to use the present technologies; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR.
4. Finally, conclusion.

MAIN FOCUS OF THE CHAPTER

16

Problems with Highly Aged So Far

Strange to say that bodily weakness will start being recognized around from the age of 75 years old to 80 years old by themselves. Legs, wrist, hands, arms, shoulder, neck, and brain will be the parts of weakened. At a glance, it seems that every bodily movement is not smooth, not in harmony, not in strength, not in active, and not in vivid. And it seems that the joints will be easily twisted, hurt, and broken. And it seems that a brain is not functioning as fast (quickly) as the decades ago. In another words, the brain responses slowly or very faintly to a stimulation. These problems will have a bad influence in a daily life to those people, which will sometimes cause a difficulty in the livings and a life.

So, as the summary of these problems mentioned above, they may be rewritten and categorized again as follows.

1. Physical Matters:

/A physical body will naturally start being weakened from around 75-80 years old.

/This phenomenon will be going further worse gradually, or seldomly it will happen suddenly. / Sometimes it will get worse and worse after the weakened.

What is more, the physical matters will be further divided into functional parts.

- a. The bones related parts.
- b. The muscles related parts.
- c. The internal organs related parts.

2. Mental Matters:

/A brain will naturally start being weakened from around 75-80 years old.

/This phenomenon will be going further worse gradually, or it will sometimes happen suddenly.

/Sometimes the one will be lost.

What is more, the mental matters will be further divided into functional parts.

- a. All over the brain (Frontal lobe, Parietal lobe, Occipital lobe, and Temporal lobe). It may seem that Temporal lobe and then Frontal lobe will be likely firstly weakened and damaged.
- b. Brain Part Focused.

Frontal lobe, Parietal lobe, Occipital lobe, and Temporal lobe.

Computer Assistance and Help

There are many kinds of AI usages in almost every field. AI is quite often used for control, an intelligent control, an intelligent system, deep learning, Cloud, a knowledge base, data base, management, a production system, statistics, sales force, environment examination, agriculture, art, livings, daily life, etc. Those usages are helpful, useful, and reliable. Because the outputs come up to us soon after commands are input by manual or by voice. And the output results will be used for practical usages of jobs, daily

life, entertainment, research, and many others. Under these situations, here looks at what the computer assistance would be able to do for the highly aged people.

In a case of Cloud, it has the following background. And those related things are already come up to the real world in a real usage.

The word of “Cloud” has come to this world many decades ago from a workstation computer company researcher. Since then, the time has passed for some decades without being paid any attention by researchers and engineers in the world. But the end of 20th century, “Cloud” comes up to this world suddenly again without any notice. Since then, Cloud related hardware and software have been fulfilled so much and so many in this world. But they are not friendly to users at all at first stage. Because the systems and the many kinds of software are not user friendly at all at the first stage. They seemed only to give a software wanted to a user platform remotely. And there are any data shares for each software. However, as the first stage, Cloud gave a good and nice impression to most of the users. Some of them are as follows (Sugiyama, 2019);

1. SAP: (Enterprise Resource Planning)
2. SOA: Service Oriented Architecture
3. SaaS: Software as a Service
4. Mash UP: Mashing Up software
5. Meta Frame: (peer to peer relation)
6. PaaS: Platform as a Service
7. HaaS: Hardware as a Service
8. IaaS: Infrastructure as a Service
9. Cloud: Cloud Computing within a particular area
10. Mini Cloud: Various Kinds of Small Cloud
11. Individual Cloud: Purposely Made Cloud for Individual
12. Fog Computing (Cloud): Internet Of Things (IoT) in Cloud
13. AI and Cloud: at present
14. VR, AR, AI, and Cloud: the coming near future

Those systems and functions mentioned above may have been generally concluded as a part of “Cloud” Computing mainly to solve problems at a company, an individual, a community, a system, a handicapped, and a society.

Nowadays, “Cloud” has become more user friendly. And as broad ways of usages, they have been offered from “Cloud” in/of a mainframe to Desktop’s, to Laptop’s, and to Smart Phone’s.

At this stage of the technologies flow, it would not be rough and over expected idea that Cloud (and Individual Cloud, Fog Computing, AI, VR, and AR) is a necessary and an important concept of the assistance and the help for highly aged people (Sugiyama, 2012. Sugiyama, 2013a. Sugiyama, 2013b. Sugiyama, 2018. Sugiyama, 2019).

SOLUTIONS AND RECOMMENDATIONS

Transform and Strengthen Weakness To Normal

As mentioned in the above section, there are two categories for assisting and helping the highly aged people, namely the physical and the mental assistance and help. In order to do these things, it is important to comprehend the vital situations (the physical matters and the mental matters) of the highly aged people.

The problems are described here again. And here studies how to transform Weakness to Normal, and to strengthen Weakness to Normal.

1. Physical Matters:

/A physical body will naturally start being weakened from around 75-80 years old.

/This phenomenon will be going further worse gradually, or seldomly it will happen suddenly. / Sometimes it will get worse and worse after the weakened.

What is more, the physical matters will be further divided into functional parts.

a. The bones related parts:

This is again divided into two. Namely, one is Prevention and the other is Medical Treatment.

i. Prevention:

There are lots of information on the medical attentions on the prevention of the bones related parts. It will be possible to use them in a daily life for the highly aged people. It can be offered by using “Cloud” and “Individual Cloud”. What the highly aged people are now doing can be predicted by the iPhone with them (necessary information: the place, the surroundings, etc.) through “Fog Computing”. By using these data, “Cloud” and “Individual Cloud” can tell the highly aged people through their iPhones about the accidents’ possibilities that will be occurring. In addition to these, it will be possible to check whether they are doing those preventions or not. What is more, these preventions in the medical attentions will be able to be offered by VR and AR through the iPhone with them. VR can offer the necessary information visually to the highly aged target people in detail or in roughly as wanted. AR can offer the necessary information virtually mixed with the reality around the highly aged target people as wanted. These matters are now possible with the present IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR.

ii Medical Treatment:

It will be possible to offer the medical treatment information to the highly aged people. It can be offered by through “Cloud” ® “Fog Computing” ® “Individual Cloud” ® The highly aged people (¬ “Individual Cloud; The related kind of Hospitals, the Nurses, and the Medical Managers” ¬ “Cloud”).

b. The muscles related parts:

This is again divided into two. Namely, one is Prevention and another is Medical Treatment.

i Prevention:

There are lots of information on the medical attentions on the prevention of the muscles related parts. It will be possible to use them in a daily life for the highly aged people. It can be offered by using “Cloud”. What the highly aged people are doing can be predicted by the cellular phone with them (necessary information: the place, the surroundings, etc.) through “Fog Computing”. By using these data, Cloud is able to tell the highly aged people through their cellular phones about the accidents possibilities that will be occurring. In addition to these, it will be possible to check whether or not they are doing those preventions. After all, these matters are now possible with the present IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR.

ii Medical Treatment:

It will be possible to offer the medical treatment information to the highly aged people. It can be offered by through “Cloud” ® “Fog Computing” ® “Individual Cloud” ® The highly aged people (→ The related kind of Hospitals, the Nurses, and the Medical Managers).

After all, these matters are now possible with the present IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR. Especially, AI, VR, and AR will role an important play for the medical treatment.

c. The internal organs:

The internal organs are able to be sensed by using the technology and the study of MEMS (Micro Electro Mechanical Systems). These are still under research and development works, but it is very important to get them in order to put them into the circulation of “Individual Cloud” ® “Cloud” ® “Individual Cloud” ® “Fog Computing” ® The highly aged people for the medical treatment.

In the same way as mentioned in the above sections, “Prevention” and “Medical Treatment” will be done by using the technologies of IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR.

2. Mental Matters:

/A brain will naturally start being weakened from around 75-80 years old.

/This phenomenon will be going further worse gradually, and or it will happen suddenly.

What is more, the mental matters may be further divided into functional parts.

a. These things will be related with all over the brain (Frontal lobe, Parietal lobe, Occipital lobe, and Temporal lobe):

It may be said that Frontal lobe and Temporal lobe will be likely firstly damaged. These are still under research works and technological developments, but it is very important to get them in order to put them into the circulation of “Individual Cloud” ® “Cloud” ® “Individual Cloud” ® “Fog Computing” ® The highly aged people for the mental matters’ enlightenment treatment.

In the same way as mentioned in the above sections, “Prevention” and “Medical Treatment” will be assisted by using the technologies of IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR. In this case, it will be very difficult to understand the situation of oneself in detail. Because the behavior of the one cannot be accurately grasped. So, Individual Cloud would play the most important role.

b. Frontal lobe (Prefrontal cortex):

The prefrontal cortex is the place where “the entity” as “human” can be determined by the groups of neuron activities. This is the place where a consciousness and a consciousness state come out. And this is the place where many abstracted matters like “love”, “desire”, “feeling”, “mind”, etc. come out in a consciousness and a consciousness state. Namely, this is the most important place and part in the brain as a humankind. These are still under research works and technological developments, but it is very important to get them in order to put them into the circulation of “Individual Cloud” ® “Cloud” ® “Individual Cloud” ® “Fog Computing” ® The highly aged people for the mental matters’ enlightenment treatment.

In the same way as mentioned in the above sections, “Prevention” and “Medical Treatment” will be assisted by using the technologies of IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR.

In the same way as mentioned in the above sections, “Prevention” and “Medical Treatment” will be assisted by using the technologies of IT; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR. In this case, it will be very difficult to understand the situation of oneself in detail. Because the behavior of the one cannot be accurately grasped. So, Individual Cloud would play the most important role.

FUTURE RESEARCH DIRECTIONS

This area is just started to study and to research.

The physical and mental situations of the highly aged people are not the same individually as known quite well. And so, it is not an easy matter to understand the directions of them to assist and help in order to get them better. In this sense, it is very important to bridge between human body and Cloud more intimately even though Cloud has got much attention to do anything for assisting and helping. So, in this Chapter, in addition of “Cloud at present”, Individual Cloud, Fog Computing, AI, VR, and AR have been added for the considerations.

However, the following things will be very important to study in the coming future:

1. To understand an individual thought of highly aged people towards life and living through Individual Cloud (and also Cloud, Fog Computing, AI, VR, AR, and MEMS).
2. To study much more about “MEMS” for the interface among human, electronics devices, and “Cloud in general”. This is the one of key issues for further advancements of this issue
3. To study about the transform mechanisms and behaviors of the highly aged people in order to moderate and to harmonize the transformations (Weakness ® Normal).
4. To study about VR (Virtual Reality), AR (Augmented Reality), and Cloud that should be used in harmonized and mixed methods in order to assist and to help the highly aged people.
5. To study about a communication method among the highly aged people by using verbal method, body language, gesture, eye contact, etc. in order to understand better.
6. To study about how to connect the related group of people in order to assist and help the highly aged people.

CONCLUSION

There have been more and more people who will live longer year by year. This phenomenon will become a global expansion of highly aged people. In this chapter, tried to find out what could be done by using the present high technologies in order not to cause troubles for those highly aged people. And studied about how to transform and how to strengthen the weakness to the normal for the highly aged people as shown below;

1. It was considered what sort of physical body and mental damages would come out to the highly aged people.
2. It was studied what sort of assistance and help would be needed for those damages.
3. It was studied what sort of assistance and help would be possible to use the present technologies; Cloud, Individual Cloud, Fog Computing, AI, VR, and AR.
4. It is understood that the present technologies are quite important for the highly aged people.

REFERENCES

- JILI. (2016). *The average life-span*. Retrieved 2019, from <https://www.jili.or.jp/lifespan/lifesecurity/oldage/2.html>
- Sugiyama, S. (2012). *Dynamics in Knowledge*. In *Knowledge Management and Drives of Innovation in Services Industries*. IGI Global.
- Sugiyama, S. (2013a). *Intangible Capital Management Method As Dynamic Knowledge Wisdom*. In *Intellectual Capital Management For Knowledge-Based Organizations*. IGI Global.
- Sugiyama, S. (2013b). *Fusion between Dry Knowledge and Wet Brain*. In *Strategic Role of Tertiary Education and Technologies for Sustainable Competitive Advantage*. IGI Global.
- Sugiyama, S. (2018). *After Cloud*. In *Hypothetical World, Deep Learning and Their Convergence With Big Data*. IGI Global. doi:10.4018/978-1-5225-3015-2.ch010
- Sugiyama, S. (2019). *Human behavior and another kind in consciousness*. IGI Global. doi:10.4018/978-1-5225-8217-5
- Tentouyobounavi. (2017). Retrieved from <https://tentouyobounavi.com/goods/148>
- Yomeishu. (2012). Retrieved 2019 from <https://www.yomeishu.co.jp/genkigenki/fature/120529/index.html>

KEY TERMS AND DEFINITIONS

AI: Artificial intelligence.

AR: Augmented reality.

Cloud (Computing): Cloud computing is a way to offer a required and a necessary information for users that are retrieved through the internet.

Fog Computing: Decentralized computing infrastructure that data, storage, IoT, and software are distributed around in the most idealized and the most efficient place between the data source and cloud.

Highly Aged People: Group of people whose ages are over 75 years old (80 years old).

Individual Cloud: The cloud that is used only for an individual person through cloud from the internet.

MEMS: Micro-electronics mechanical systems.

MR: Mixed reality.

VR: Virtual reality.

Gene Editing Technology and Ethical Issues

Barbara Jane Holland

 <https://orcid.org/0000-0003-3729-0147>

Brooklyn Public Library, USA

INTRODUCTION

Technologies for making accurate additions, deletions, and alterations to DNA—have generated interest across the world because of the promise it holds to improve human health. For example, genome editing is being tested in clinical trials to engineer immune cells to target cancerous tumor cells and to make cells more resistant to HIV. Genome editing could also be used to develop new treatments for devastating genetic diseases like Huntington’s disease, sickle cell anemia, immune deficiencies, muscular dystrophy, and cystic fibrosis. Technology as tool view is only part of the story about the complex relationships between technology, society, and ethics. This is why technologies, are also complex social phenomena. Some technologies are encouraged by society through social demand or public funding. (Sandler R.L. 2014) This is often the case with medical technologies, while some technologies are opposed or rejected by society (or at least by some members of society)

Important inquiries have been raised about human genome editing and balancing potential benefits with the risk of unintended harms and using this type of technology while incorporating societal values into clinical applications and policy decisions. Furthermore, respecting the inevitable differences across nations and cultures that will shape diverse perspectives regarding these technologies. The recent development and growing use of the CRISPR/Cas9 system have resulted in an explosion of research in genome editing. Clinical trials are testing how these technologies can be used to improve health.

BACKGROUND

Five years ago University of California, Berkeley, biochemist Jennifer Doudna, and colleagues unveiled the gene-editing tool CRISPR (Science 2012). Doudna’s lab, in fact, was engaged in a patent fight with the Harvard- and MIT-affiliated Broad Institute.

The patent case turns on the question of which researchers at the two institutions conceived of the most important CRISPR applications first. UC says it’s a team at its Berkeley lab headed by Jennifer Doudna, with the collaboration of Emmanuelle Charpentier, now of Germany’s Helmholtz Centre for Infection Research. Broad’s claim is based on the work of its researcher Feng Zhang.

The patent office changed its rules to a “first-to-file” basis from the old “first-to-invent,” in 2013, making the U.S. the last major country to do so. The idea was to end disputes that turned on minute interpretations of lab records or personal notes, with costs that arguably disadvantaged small inventors.

Last February 2017, the U.S. Patent Trial and Appeal Board ruled that although a team led by UC Berkeley structural biologist Jennifer Doudna had first laid claim to the use of CRISPR to cut DNA in a

DOI: 10.4018/978-1-7998-3479-3.ch136

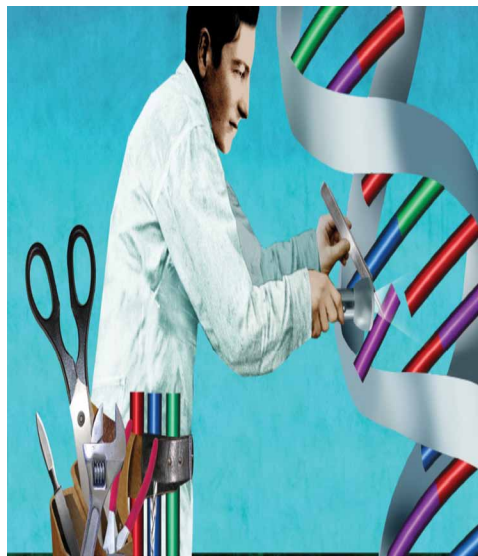
test tube, the use of the method on human cells by molecular biologist Feng Zhang's team at the Broad was still an advance.

However, a decision from the European Patent Office (EPO) in January 2018 revoked the patent from the Broad Institute of MIT and Harvard University because the Broad did not meet EPO requirements to establish that its researchers were the first to use CRISPR in eukaryotes. The patent granted to the Broad for fundamental aspects of the technology is one of several of its patents facing opposition in Europe.

Genome editing (also called gene editing) is a group of technologies that give scientists the ability to change an organism's DNA. These technologies can add, alter, or genetic material at particular locations in the genome. Several approaches to genome editing have been developed. Once the DNA is cut, researchers use the cell's own DNA repair machinery to add or delete pieces of genetic material, or altering the DNA by replacing an existing segment with a customized DNA sequence.

Genome editing is of great interest in the prevention and treatment of human diseases. It is being explored in research on a wide variety of diseases. It also holds promise for the treatment and prevention of a more complex disease, such as human immunodeficiency virus (HIV) infection.

Figure 1.



Clustered Regularly Interspaced Short Palindromic Repeats

One of most recent gene editing technology getting a lot of attention is CRISPR-Cas9. The CRISPR-Cas9 system has generated a lot of enthusiasm in the scientific community because it is faster, cheaper, more accurate, and more efficient than other existing genome editing methods.

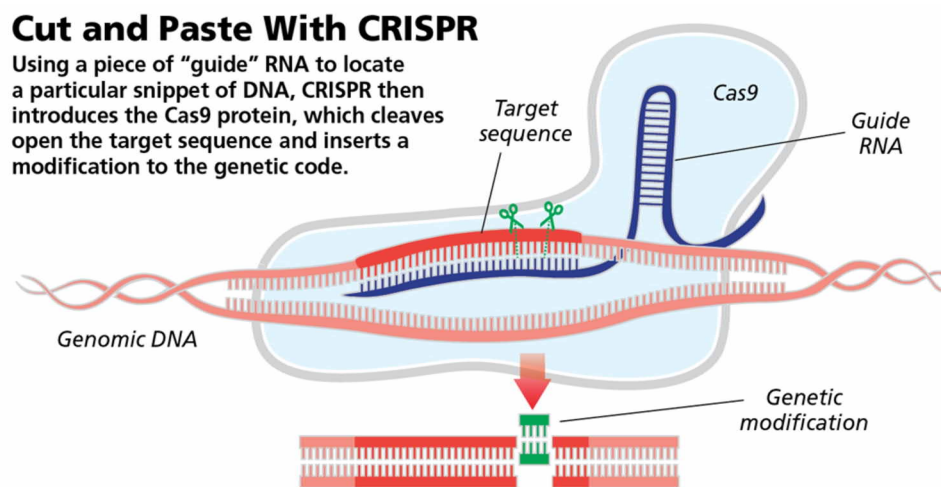
CRISPR technology is a simple yet powerful tool for editing genomes. It allows researchers to easily alter DNA sequences and modify gene function. Its many potential applications include correcting genetic defects, treating and preventing the spread of diseases and improving crops. However, its promise also raises ethical concern (Vidyasagar A 2018).

CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats, which is a sequence of base pairs found in the DNA of bacteria that have this feature. Some bacteria and archaea use CRISPR

for adaptive immunity, in order to incorporate bits of DNA from invading viruses into segments of their own DNA in order to target their immune system at those sequences.

CRISPR-Cas9 was adapted from a naturally occurring genome editing system in bacteria. The bacteria capture snippets of DNA from invading viruses and use them to create DNA segments known as CRISPR arrays. The CRISPR arrays allow the bacteria to “remember” the viruses (or closely related ones). If the viruses attack again, the bacteria produce RNA segments from the CRISPR arrays to target the viruses’ DNA. The bacteria then use Cas9 or a similar enzyme to cut the DNA apart, which disables the virus.

Figure 2.



CRISPR/Cas is an adaptive immune system (reviewed in Hryhorowicz et al. 2017) found in many bacteria and archaea, which enables effective defense against the invasion of bacteriophages or viruses. This immune system allows prokaryotes to “memorize” foreign DNA by incorporating its fragments into CRISPR arrays and ensures fast response to another infection in the future (Barrangou et al. 2007).

The CRISPR-Cas9 system works quasi like in the lab. Researchers create a small piece of RNA with a short “guide” sequence that attaches to a specific target sequence of DNA in a genome. The RNA also binds to the Cas9 enzyme. likewise, with bacteria, the modified RNA is used to recognize the DNA sequence, and the Cas9 enzyme cuts the DNA at the targeted location.

The CRISPR/Cas immune system performs its function in three general stages: adaptation, expression, and interference. During the first stage, short fragments of viral or phage DNA are incorporated into the CRISPR alignment. The integration of the new viral/phage DNA sequences is followed by duplication of a repeat, which in this way is forming a new spacer-repeat unit (Zych A O 2017). Spacer precursors called proto-spacers are selected from invading DNA depending on the recognition of neighboring protospacer-adjacent motifs (PAM). PAM sequences are typically several nucleotides long and vary among different variants of the CRISPR/Cas system (Makarova et al. 2011). The arrangement of spacers within the CRISPR alignment corresponds to the sequence of invasion events. In the next stage, CRISPR alignment is transcribed and primary transcript pre-crRNA is produced, which then is processed to mature CRISPR RNA (crRNA) by RNase III. Depending on the CRISPR/Cas system class, this process can be mediated either by multiprotein CRISPR ribonucleoprotein complex or a single protein (Zych, A.O).

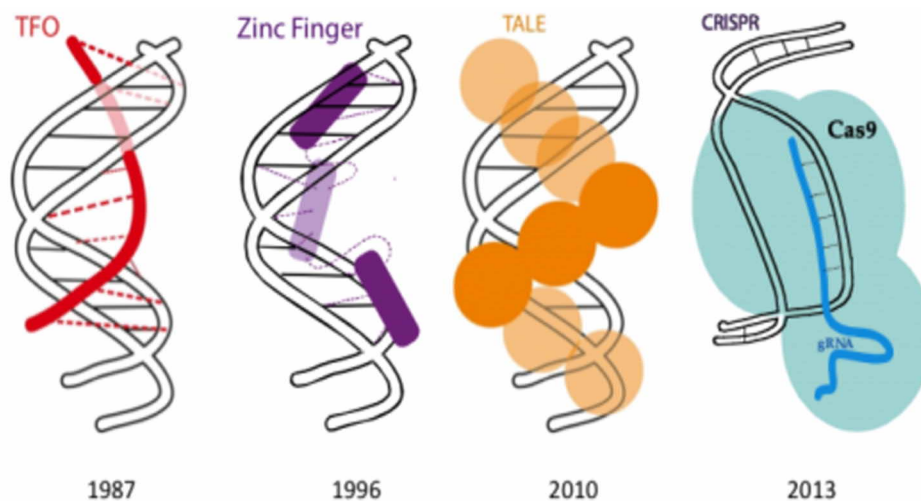
Despite the variety of the CRISPR/Cas systems in nature, the most commonly used type adapted to genome editing is class 2 type II CRISPR/Cas9. CRISPR/Cas9 requires two short RNA sequences: crRNA and trans activating crRNA (tracrRNA) to recognize and cleave foreign DNA sequences. During the action of Cas9, the crRNA hybridizes with the tracrRNA forming duplex crRNA:tracrRNA, which in the next step associates with Cas9 (Zych, A.O). Despite the high efficiency, feasibility, and simplicity of target design CRISPR/Cas9 technique faces important complications. The essential limitation of this method is mutations at sites with similar but not identical homology to the target sites (Cradick et al. 2013).

There are several clinical trials in China investigating the efficacy of CRISPR/Cas9-mediated PD-1 knock-out T cells in patients with malignancies such as advanced esophageal carcinoma, muscle-invasive bladder cancer, metastatic renal cell carcinoma and others.

Gene-Editing Technology & Tools

Genome-editing tools all work in a similar fashion. They “target” specific DNA sequences for individual genes or non-coding regions by engineering certain proteins or protein-RNA complexes that can then recognize and bind the sequences and generate single-strand or double-strand DNA breaks. For example, a Cas9 protein along with a CRISPR “guide RNA” can find a target gene among the thousands of genes in a cell’s genome and cleave both DNA strands at the target site. Genome editing is already accelerating the pace of many areas of biology as researchers use genome-editing methods to more quickly and cheaply study the function of genes in model organisms and economically important species such as crops, livestock, and energy feedstock.

Figure 3.



Emerging Technologies and Ethical Concerns

Ethical technology tends to focus on the evaluation of emerging technologies and what might be problematic. The presumption is that new technologies are favorable. The question is not “why should we pursue or permit this new technology, but rather” are there any good reasons not to develop it”.

A prominent issue in the ethics of emerging technology is what role people's moral or religious concerns ought to play in developing policy and regulation.

When the potential of the CRISPR-Cas9 system was initially realized, ethical concerns were raised about the possibility of creating permanent and inheritable changes in the genome of human gametes and embryos. As a response to this development, a call for a moratorium was made for research involving genome modification of the human germline, or more specifically for the clinical applications of such technologies.

Most of the ethical discussions related to genome editing center around human germline editing. This is because changes made in the germline would be passed down to future generations.

Figure 4.



Bioethicists and researchers believe that human genome editing for reproductive purposes should be halted at this time, but continue studies that would make gene therapy safe and effective.

The first known attempt at creating genetically modified human embryos in the United States has been carried out by a team of researchers led by Shoukhrat Mitalipov of Oregon Health & Science University in Portland, Oregon in 2017. The effort, led by Shoukhrat Mitalipov of Oregon Health and Science University, involved changing the DNA of a large number of one-cell embryos with the gene-editing technique CRISPR, according to people familiar with the scientific results. Embryos were allowed to develop for more than a few days—and the researchers said there was never any intention of implanting them into a womb—the experiments are considered by many a milestone on what may prove to be an inevitable journey toward the birth of the first genetically modified humans.

Mitalipov and his collaborators showed they could bump up the efficiency of human embryo editing by inserting the CRISPR machinery earlier in development than previous experiments (Nature 548). When they combined healthy eggs with sperm bearing a disease-causing mutation and immediately added CRISPR, they found that 72% of the resulting embryos were free of the mutation—rather than the expected 50% that would have avoided inheriting the harmful gene anyway.

Although the researchers inserted short strands of DNA as templates for repair, the cells didn't seem to pick them up; those specific sequences were absent from the embryos. The team concluded that the cells must have relied instead on the non-mutated sequence in the egg donor's DNA when making the repairs.

Developmental biologist Maria Jasin of Memorial Sloan Kettering Cancer Center in New York City and Columbia University stem cell biologist Dieter Egli challenged that interpretation. The authors, including well-known CRISPR researcher and Harvard University geneticist George Church, say that the Nature paper goes against conventional wisdom about how embryos are organized early in development. Right after an egg is fertilized, the DNA from the sperm and the egg aren't believed to be in close enough proximity to interact or share genes, they explain.

Raising Ethical Questions

Bioethicists and researchers generally believe that human genome editing for reproductive purposes should not be attempted at this time, but that studies that would make gene therapy safe and effective should continue. Most agree that it is imperative to have continuing public deliberation and debate to allow the public to decide whether or not germline editing should be permissible.

The Hugo Committee

The Human Genome Organization (HUGO) has a long tradition of advocacy for 'benefit sharing' to realize societal as well as economic opportunities.

The HUGO Committee on Ethics, Law, and Society recently stated that 'genomic solidarity' ideally supports collaborations between individuals, communities, and populations, with research communities and industry. Significantly, benefit sharing and genomic solidarity work together through an idea of the public good. The public good is valued distributively, i.e., to each and every person that value is secured through equality of rights.

The idea of genomic solidarity is likely to be challenged, for it confronts the ingrained idea of public goods as something which hinders the benefits of economies of exclusion and rivalry. In that classical estimation, goods are 'public' depending on whether a private investment has any interest in them; in other words, if a good is profitable, then it is economically wasteful to consider it a public one. Paul Samuelson originally called it collective consumption goods (what later became known as 'public goods'). It leads to a particular view of goods that can be applied to the human genome.

Ethical Concerns and Gene Editing

Ethicists become concerned when germ cells are the target of CRISPR. Any changes in the germ cells can be potentially passed down for future generations, essentially introducing those changes into the human population.

Critics, however, worry that if the technology is developed to treat diseases it is a slippery slope to treating unwanted characteristics that are not actually diseases. Parents wanting stronger, smarter, taller children, for example, might be tempted by such technology.

Human intervention has long been discussed and debated in religious and spiritual terms. Christian traditions in the contemporary West, have had the most influence, on what today considered a most religiously diverse and secular culture. These ideas are expressed in the debate about which tasks in improving or stewarding nature are the domain and obligations of humans and which are left up to God.

With the Jewish tradition, there is an explicit obligation to build and develop the world that will be beneficial to human beings. Such improvements are viewed as a positive collaboration between God and humans and not as an interference with creation. (Steinberg, 2006). Muslim scholars and Jurists often respond to new technologies once they have already been introduced into Muslim societies. Assessment of such technologies is often taken in a piecemeal fashion. Technologies are analyzed on a case by case basis.

CRISPR-Cas9 forces scientists, ethicists and people of faith alike to consider where to draw the line on gene editing. There is a deep concern that couples will someday seek to create designer babies through gene editing.

Figure 5.



Furthermore, there is concern that genome editing will only be accessible to the wealthy and will increase existing disparities with access to health care and other interventions. Some worry that taken to its extreme, germline editing could create classes of individuals defined by the quality of their engineered genome.

Many people have moral and religious objections to the use of human embryos for research. Federal funds cannot be used for any research that creates or destroys embryos. In addition, NIH does not fund any use of gene editing in human embryos.

While new scientific advances have made gene editing easier and open possibilities for improved treatment and prevention of genetic diseases, the technology has risks, including the unintentional alteration of other genes, and ethical concerns, such as the introduction of mutations that will impact all future progeny.

Drawbacks and Challenges

Germline editing poses many important issues, including:

- the risks of inaccurate editing (such as off-target mutations) and incomplete editing of the cells of early-stage embryos (mosaicism);
- the difficulty of predicting harmful effects that genetic changes may have under the wide range of circumstances experienced by the human population, including interactions with other genetic variants and with the environment;
- the fact that, once introduced into the human population, genetic alterations would be difficult to remove and would not remain within any single community or country;
- the possibility that permanent genetic ‘enhancements’ to subsets of the population could exacerbate social inequities or be used coercively
- the moral and ethical considerations in purposefully altering human evolution using this technology.

Solutions & Recommendations

Possible Strategies for Working On The Immune System

- Instead of delivering CRISPR/Cas9 into the body, you take cells out of the body, use CRISPR to edit their genes in a lab, and return Cas9-free cells. This is the strategy pursued by CRISPR Therapeutics for the inherited blood disorder thalassemia and in various trials using CRISPR to modify immune cells to attack cancer
- Only use CRISPR in places the immune system cannot reach such as the eye. The immune system can’t really attack invaders there.
- It may be possible to redesign Cas9 to hide it from the immune system or to find other bacterial proteins that can do the job of Cas9 without provoking the immune response. Many different bacteria have CRISPR systems
- Researchers should use existing regulatory processes for human gene therapy to oversee somatic human genome editing research and uses
- Limit clinical trials or therapies to treatment and prevention of disease or disability at this time
- Evaluate safety and efficacy in the context of risks and benefits of intended use
- Researchers should permit clinical research trials only for compelling purposes of treating or preventing serious disease or disabilities, and only if there is a stringent oversight system able to limit uses to specified criteria
- Ongoing reassessment and public participation should precede any heritable germline editing Enhancement
- Scientists should not proceed at this time with human genome editing for purposes other than treatment or prevention of disease and disability
- Scientists should encourage public discussion and policy debate with respect to somatic human genome editing for uses other than treatment or prevention of disease and disability

FUTURE RESEARCH DIRECTIONS

Recently a nonpeer review study posted on bioRxiv 2018 reported that many people have existing immune proteins and cells primed to target the Cas9 proteins included in CRISPR complexes. That means those patients might be immune to CRISPR-based therapies or vulnerable to dangerous side effects — the latter being especially concerning as CRISPR treatments move closer to clinical trials. (Carsten C

et; al). Instead of successfully modifying the genome, when used therapeutically, the tool could trigger an adaptive immune response. The human body could treat Cas9 proteins as foreign bodies. This may trigger immune responses that would block gene therapy and pose a greater threat to the patient's health than the original disease they're treating for.

However, researchers not involved with the study said its findings, if substantiated, could be bypassed. Many of the first planned CRISPR clinical trials, for example, involve removing cells from patients, fixing their DNA, and then returning them to patients. In that case, it's possible that there will be few or no CRISPR proteins remaining for the immune system to detect. Researchers found that 79% of study participants produced antibodies against Cas9 derived from *S. aureus* and 65% from *S. pyogenes*.

As a safety precaution added to the CRISPR therapy protocol, at-risk patients could be screened before undergoing gene-editing treatment. A way around the immune system obstacle would be to develop other Cas9 enzymes from other bacteria that don't trigger an immune response.

CRISPR Therapeutics

CRISPR Therapeutics, based in Cambridge, Massachusetts, became the first company to ask permission from European regulators to begin a trial in 2018. The company will use the gene-editing technology on patients with beta thalassemia, an inherited blood disorder.

Daniel Anderson of the Massachusetts Institute of Technology, who has studied the delivery of CRISPR therapies, would like to conduct experiments to learn more. Likewise, Dr. Matthew Porteus and colleagues of Stanford, University is at work on a CRISPR-based therapy for sickle cell disease.

CRISPR Therapeutics is also planning to request approval from the U.S. Food and Drug Administration in the first half of 2018 to begin a CRISPR trial for sickle-cell disease. A Researcher Team led by Matthew Porteus, an associate professor of pediatrics at Stanford University School of Medicine are moving a CRISPR treatment for sickle-cell to clinical trials and plans to seek authorization from the FDA in 2018 to begin trials the following year

The center of the issue is a protein called Cas9, which is used as part of CRISPR-Cas9 to target and cut out specific sections of DNA. Without Cas9, CRISPR wouldn't work, but it's this specific protein that our bodies have learned to fight. Cas9 is typically found in harmful bacteria like *Staphylococcus aureus* and *Streptococcus pyogenes*, which cause staph and strep infections, respectively, so it's usually a good thing that our bodies fight the protein.

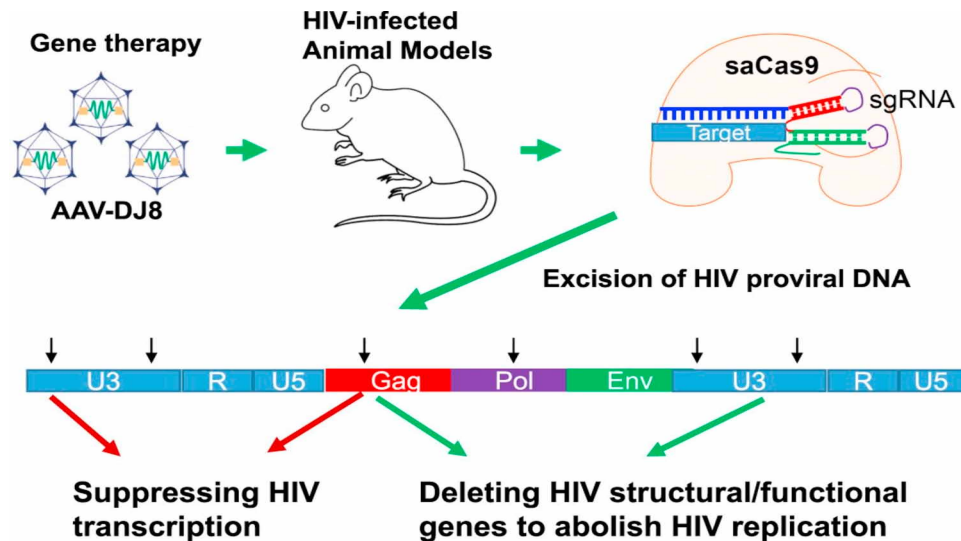
However, this does present a hurdle for researchers learning how to use CRISPR in human trials. There are already some tricks that researchers are using to circumvent the immune systems, like only using CRISPR outside of the body or in places that immune cells can't reach, but ultimately having to abandon Cas9 for a different protein that doesn't set off the bodies' defenses. CRISPR researchers could be facing a serious setback if this holds true.

CRISPR and AIDS

To eliminate the integrated HIV-1 genome, (Kaminski, R. *et al* 2016) used the Cas9/guide RNA (gRNA) system, in single and multiple configurations. The key components of CRISPR/Cas9 are guided RNAs (gRNAs) which determine specific sequence targeting of DNAs. Highly specific targets within the HIV-1 LTR U3 region were identified and efficiently edited by Cas9/gRNA, inactivating viral gene expression and replication in latently infected microglial, promonocytic, and T cells. Cas9/gRNAs caused neither genotoxicity nor off-target editing to the host cells, and completely excised a 9,709-bp fragment of

integrated proviral DNA that spanned from its 5' to 3' LTR. Furthermore, the presence of multiplex gRNAs within Cas9-expressing cells prevented HIV-1 infection. Their results suggest that Cas9/gRNA can be engineered to provide a specific, efficacious prophylactic and therapeutic approach to AIDS.

Figure 6.



The ultimate cure for HIV/AIDS will be the removal or disruption of integrated HIV provirus in latently infected cells or the elimination of these latent cells completely. (Zaohua H & Nair 2017)

FUTURE RESEARCH TRENDS

CRISPR Therapeutics AG & PESTEL Analysis

CRISPR Therapeutics AG PESTEL Analysis is a strategic tool to analyze the macro environment of the organization.

PESTEL stands for - Political, Economic, Social, Technological, Environmental & Legal factors that impact the macro environment of CRISPR Therapeutics AG.

Changes in the macro-environment factors can have a direct impact on not only the CRISPR Therapeutics AG but also can impact other players in the Biotechnology.

For example, an Industry may be highly profitable with a strong growth trajectory but it won't be any good for CRISPR Therapeutics AG if it is situated in the unstable political environment.

Two major concerns

Gene-edited products are subjected to the GMO's strict directive

1. Increased checks and controls would make it even more expensive for European academics and small companies to develop products for the market.
2. . the public could reject these products if they are considered GMOs, as was the case for genetically modified foods. CRISPR gene editing technology could play a key role in the fight against new

Table 1. CRISPR Therapeutics AG PESTEL & Environment Analysis

Political Factors that Impact CRISPR Therapeutics AG	● Political stability and importance of Biotechnology sector in the country's economy. ● Risk of military invasion ● Level of corruption - especially levels of regulation in Healthcare sector. ● Bureaucracy and interference in Biotechnology industry by government. ● Legal framework for contract enforcement ● Intellectual property protection ● Trade regulations & tariffs related to Healthcare ● Favored trading partners ● Anti-trust laws related to Biotechnology
Economic Factors that Impact CRISPR Therapeutics AG	Economic factors that CRISPR Therapeutics AG should consider while conducting PESTEL analysis are - ● Type of economic system in countries of operation – what type of economic system there is and how stable it is. ● Government intervention in the free market and related Healthcare ● Exchange rates & stability of host country currency. ● Efficiency of financial markets – Does CRISPR Therapeutics AG needs to raise capital in local market? ● Infrastructure quality in Biotechnology industry ● Comparative advantages of host country and Healthcare sector in the particular country. ● Skill level of workforce in Biotechnology industry. ● Education level in the economy
Social Factors that Impact CRISPR Therapeutics AG	Social factors that leadership of CRISPR Therapeutics AG should analyze for PESTEL analysis are - ● Demographics and skill level of the population ● Class structure, hierarchy and power structure in the society. ● Education level as well as education standard in the CRISPR Therapeutics AG 's industry ● Culture (gender roles, social conventions etc.) ● Entrepreneurial spirit and broader nature of the society. Some societies encourage entrepreneurship while some don't. ● Attitudes (health, environmental consciousness, etc.) ● Leisure interests
Technological Factors that Impact CRISPR Therapeutics AG	A firm should not only do technological analysis of the industry but also the speed at which technology disrupts that industry. Slow speed will give more time while fast speed of technological disruption may give a firm little time to cope and be profitable. Technology analysis involves understanding the following impacts - ● Recent technological developments by CRISPR Therapeutics AG competitors ● Technology's impact on product offering ● Impact on cost structure in Biotechnology industry ● Impact on value chain structure in Healthcare sector ● Rate of technological diffusion
Environmental Factors that Impact CRISPR Therapeutics AG	Before entering new markets or starting a new business in existing market the firm should carefully evaluate the environmental standards that are required to operate in those markets. Some of the environmental factors that a firm should consider beforehand are - ● Weather ● Climate change ● Laws regulating environment pollution ● Air and water pollution regulations in Biotechnology industry ● Recycling ● Waste management in Healthcare sector ● Attitudes toward "green" or ecological products ● Endangered species ● Attitudes toward and support for renewable energy
Legal Factors that Impact CRISPR Therapeutics AG	● Anti-trust law in Biotechnology industry and overall in the country. ● Discrimination law ● Copyright, patents / Intellectual property law ● Consumer protection and e-commerce ● Employment law ● Health and safety law ● Data Protection

diseases that arrive in Europe due to globalization and climate change, so the prospect of slowing their progress even further is worrying.

What's more, it appears that there is nothing stopping individual EU member states from making their own rules for gene editing. This raises the possibility that there could be a patchwork of different rules across the continent, which could undermine the single market

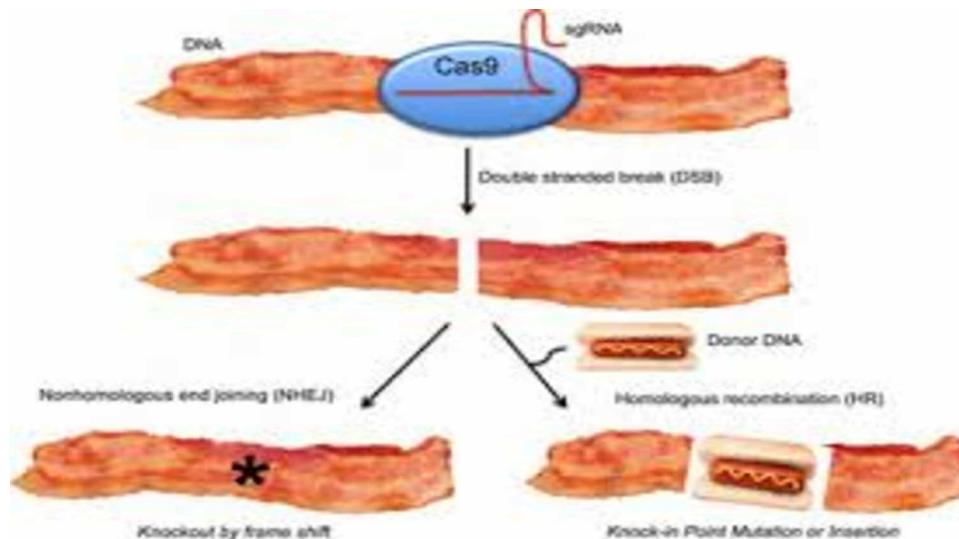
CONCLUSION

The development of CRISPR–Cas9 as an effective genome-editing tool is under perusal because it brings with it the possibility that scientists could make permanent modifications to the human germline.

Scientists and medical personnel have been publishing their research experiments since the onset of CRISPR. The single, major giant that supports biomedical sciences, including genomics, is 'the public' or, realistically, a constellation of all public. The concept of the public good has been articulated as solidarity, a community that finds it worthy to protect the shared interests of research. Regulators, funders, scientists, and editors need to continue working together to define the details of the path forward for germline genome editing so that the valuable resources and tools now at our disposal are used with good judgment.

The Beijing-based researchers have already used Crispr to take a version of UPC1 from mice and inserted it into pig embryos, which they then implanted into 13 female pigs. Three of them became pregnant and gave birth to 12 piglets, and subsequently, low-fat bacon was created.

Figure 7.



Public education and engagement are crucial in the process of assessing and applying societal values to the risks and benefits of genome editing technologies and the ethical dimensions they involve. There should be active participation and input by the public, along with an ongoing reassessment of both health and societal benefits and risks, before moving clinical trials ahead.

While researchers keep refining the gene editing tool, CRISPR is becoming very popular among DIY scientists and biohackers. Some believe that the relatively simple methods that this technique requires

might help democratize science and bring it closer to people outside the lab. However, some recent cases of biohackers injecting themselves with experimental treatments have alarmed the public and it remains to be seen how these uses will be regulated

REFERENCES

- Aparna, V. (2018). What is Crispr. *Live Science*. <https://www.livescience.com/58790-crispr-explained.html>
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819), 1709–1712. doi:10.1126/science.1138140 PMID:17379808
- Braverman, I. (2017). *Gene Editing, Law, and the Environment: Life Beyond the Human*. Routledge. doi:10.4324/9781315168418
- Capps, B., Chadwick, R., Joly, Y., Mulvihill, J. J., Lysaght, T., & Zwart, H. (2017). Falling giants and the rise of gene editing: Ethics, private interests, and the public good. *Human Genomics*, 11(20), 20. Advance online publication. doi:10.1186/40246-017-0116-4 PMID:28851444
- Charlesworth, C. (2018, January 5). *Identification of Pre-Existing Adaptive Immunity to Cas9 Proteins in Humans*. Retrieved June 6, 2018, from <https://www.biorxiv.org/content/biorxiv/early/2018/01/05/243345.full.pdf>
- Doudna, J. (2015, December 22). *Genome-editing revolution: My whirlwind year with CRISPR*. Retrieved February 13, 2017, from <http://www.nature.com/news/genome-editing-revolution-my-whirlwind-year-with-crispr-1.19063>
- Doudna, J., & Sternberg, S. (2017). *A crack in Creation, Gene Editing and unthinkable power to control evolution*. Houghton Mifflin Harcourt.
- Green, R. (2007). *Babies by Design: The Ethics of Genetic Choice*. Yale Press.
- Hryhorowicz, M. (2016, October 3). *CRISPR/Cas9 Immune System as a Tool for Genome Engineering*. Retrieved July 5, 2017, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5434172>
- Human Gene Editing Initiative. (2017). Retrieved Oct. & Nov. 2017, from National Academies of Science Engineering and Medicine website: <https://nationalacademies.org/gene-editing/index.htm>
- Human Genome Organization. (2017). *HUGO Committee on Ethics, Law, and Society*. Retrieved from <http://www.hugo-international.org/HUGO-CELS>
- International Bioethics Committee. (Rep. No. SHS/YES/IBC-22/15/2 REV.2). (2015). <https://unesdoc.unesco.org/images/0023/002332/233258E.pdf>
- Kaminski, R., Chen, Y., Fischer, T., Tedaldi, E., Napoli, A., Zhang, Y., Karn, J., Hu, W., & Khalili, K. (2016). Elimination of HIV-1 Genomes from Human T-lymphoid Cells by CRISPR/Cas9 Gene Editing. *Scientific Reports*, 6(1), 22555. doi:10.1038/rep22555 PMID:26939770
- Knoppers, B. (1999). Sovereignty and sharing. In T. Caulfield & B. Williams-Jones (Eds.), *The commercialization of genetic research: ethical, legal and policy issues* (pp. 1–11). Kluwer Academic/Plenum Publishers. doi:10.1007/978-1-4615-4713-6_1

Ledford, H. (2017, August 03). CRISPR fixes disease gene in viable human embryos. *Nature*, 548(7665), 13–14. doi:10.1038/nature.2017.22382 PMID:28770860

Ma, H., Marti-Gutierrez, N., Park, S., Wu, J., Lee, Y., Suzuki, K., Koski, A., Ji, D., Hayama, T., Ahmed, R., Darby, H., Van Dyken, C., Li, Y., Kang, E., Park, A.-R., Kim, D., Kim, S.-T., Gong, J., Gu, Y., ... Mitalipov, S. (2017). Correction of a pathogenic gene mutation in human embryos. *Nature*, 548(7668), 413–419. doi:10.1038/nature23305 PMID:28783728

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., van der Oost, J., & Koonin, E. V. (2011). Evolution and classification of the CRISPR-Cas systems. *Nature Reviews. Microbiology*, 9(6), 467–477. doi:10.1038/nrmicro2577 PMID:21552286

National Academies of Sciences, Engineering, and Medicine. (2017). *Human Genome Editing: Science, Ethics, and Governance*. Washington, DC: The National Academies Press. . doi:10.17226/24623

Orman, K. E. (2000). Human Germline Genome Editing. *American Journal of Human Genetics*, 101(2), 167–176. doi:10.1016/j.ajhg.2017.06.012

Qasim, W., Zhan, H., & Samarasinghe, S. (2017). Molecular remission of infant B-ALL after infusion of universal TALEN gene-edited CAR T cells. *Sci Transl Med*, 9.

Raquib, A. (2015). *Islamic Ethics of Technology: An Objectives' (Masjid) Approach*. The Other Press.

Samuelson, P. (1954). The pure theory of public expenditure. *The Pure Theory of Public Expenditure. The Review of Economics and Statistics*, 36(4).

Sandler, R. L. (Ed.). (2014). *Ethics and Emerging Technologies*. London: Palgrave MacMillan.

Schaefer, K. A., Wu, W.-H., Colgan, D. F., Tsang, S. H., Bassuk, A. G., & Mahajan, V. B. (2017, May 30). Unexpected mutations after CRISPR-Cas9 editing in vivo. *Nature Methods*, 14(6), 547–548. doi:10.1038/nmeth.4293 PMID:28557981

Take stock of research ethics in human genome editing. (2017, September 20). Retrieved October 2, 2017, from <https://www.nature.com/news/take-stock-of-research-ethics-in-human-genome-editing-1.22632>

Terns, R. M., & Terns, M. P. (2013). The RNA- and DNA-targeting CRISPR–Cas immune systems of *Pyrococcus furiosus*. *Biochemical Society Transactions*, 41(6), 1416–1421. doi:10.1042/BST20130056 PMID:24256230

Thomas, C. (2013, November 1). Fine Eli, et:al, CRISPR/Cas9 systems targeting β -globin *CCR5* genes have substantial off-target activity. *Nucleic Acids Research*, 41(20), 9584–9592. doi:10.1093/nar/gkt714 PMID:23939622

Zaohua, H., & Nair Madhavan, A. (2017). CRISPR/Cas9 guidance RNA screen platform for HIV provirus disruption and HIV/AIDS gene therapy in astrocytes. *Scientific Reports*, 7(1), 5955. Advance online publication. doi:10.1038/41598-017-06269-x PMID:28729655

Zych, A.O., Bajor, M. & Zagodzdzon, R. (2018). *Arch. Application of Genome Editing Techniques in Immunology*. doi:10.1007/00005-018-0504-z

KEY TERMS AND DEFINITIONS

CRISPR/Cas9-CRISPR-Cas Systems: Prokaryotic immune systems that confer resistance to foreign genetic elements such as plasmids and phages. CRISPR-Cas systems have been exploited for targeted genome editing.

Ethical Technology: Tends to focus on the evaluation of emerging technologies and what might be problematic.

Genome Editing: Genome editing or genome editing with engineered nucleases (GEEN) is a type of genetic engineering in which DNA is inserted, deleted or replaced in the genome of a living organism using engineered nucleases, or “molecular scissors.”

Germline Editing: Human germline engineering is the process by which the genome of an individual is edited in such a way that the change is heritable.

Human Germline Engineering: The process by which the genome of an individual is edited in such a way that the change is heritable.

Mosaicism: A condition in which cells within the same person have a different genetic makeup.

About the Contributors

Mehdi Khosrow-Pour, D.B.A., received his Doctorate in Business Administration from the Nova Southeastern University (Florida, USA). Dr. Khosrow-Pour taught undergraduate and graduate information system courses at the Pennsylvania State University – Harrisburg for almost 20 years. He is currently Executive Editor at IGI Global (www.igi-global.com). He also serves as Executive Director of the Information Resources Management Association (IRMA) (www.irma-international.org) and Executive Director of the World Forgotten Children Foundation (www.worldforgottenchildren.org). He is the author/editor of more than 100 books in information technology management. He is also currently the Editor-in-Chief of the *International Journal of E-Politics (IJEP)* and the *International Journal of Semiotics and Visual Rhetoric (IJSVR)*, and is also the founding Editor-in-Chief of the *Information Resources Management Journal (IRMJ)*, *Journal of Electronic Commerce in Organizations (JECO)*, *Journal of Cases on Information Technology (JCIT)*, and the *Journal of Information Technology Research (JITR)*, and has authored more than 50 articles published in various conference proceedings and scholarly journals.

* * *

Balikis Animasaun Abdul-Jeleel holds an OND from the Polytechnic of Ibadan, a B.Sc in Computer Science from the University of Ibadan and MSc in Information Science from the Africa Regional Centre for Information Science, University of Ibadan. She is presently a service desk and procurement analyst at the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria.

C. Abinaya is a student at Thiagarajar College of Engineering. Her research area include Network and Information Security.

Hisham Abusaada is a Professor of Architecture and Urban Design at the Housing and Building National Research Centre (HBRC), Egypt. During 1995-2015, he taught urban design course at universities in Saudi Arabia and Egypt. He is the author and co-authors of several scientific manuscripts and books in both Arabic and English published in scientific journals and conference proceedings. His area of interest is in urban design between theory and practice.

Olga Acosta is a PhD in Computer Sciences at the UNAM (México) in 2013. CTO of Singularityta SpA (2019).

About the Contributors

César Aguilar is a PhD in Linguistics at the UNAM (Mexico) in 2009. Assistant Professor at the Pontificia Universidad Católica de Chile.

Rukhsana Ahmed (PhD, Ohio University) is Associate Professor and Chair in the Department of Communication and Research Associate at the Center for the Elimination of Minority Health Disparities, University at Albany, State University of New York, USA. Her primary research interests lie at the intersections of health, organizational, interpersonal, and intercultural communication. In particular, her research embraces the cross/trans/inter-disciplinary nature of communication scholarship, with specific focus on exploring the role of communication processes in improving health outcomes among marginalized communities and promoting broader organizational and social changes and innovations in local, national, and international health contexts. Her collaborative and partnership research projects have been funded by Federal Government Agencies and Not-For-Profit Organization, and she has presented and published top papers and an award winning book at the intersection of health communication, service quality, culture, and media in respected international interdisciplinary and transdisciplinary conferences and peer-reviewed publication outlets. She is the Chief Editor of Health Communication, specialty section of *Frontiers in Communication*.

Kolawole Sunday Akinsunmbo is an information technology consultant, programming instructor, management training facilitator, startup management consultant and professional B2B deal broker. I have worked with brands across Nigeria as an I.T and Management consultant to deliver ease of business process, employee retention and develop effective and efficient organisational modus operandi. Kolawole is a Certified Web Programmer and Building Android and iOS Apps with Dreamweaver CC and PhoneGap (certified by Lynda.com). He a graduate member of the Nigeria Institute of Management, Certified Entrepreneurial Professional by Osun State University where I also obtained a Bachelor of Science (B.Sc) Hons in Business Administration.

Rashida Mebude Aligiwa is an O365 Support Engineer at Tek Experts. She holds a B.Sc in Physics from the Ahmadu Bello University, Zaria and a Master of Information Science (M.Inf.Sc) degree of the Africa Regional Centre for Information Science (ARCIS), University of Ibadan, Nigeria.

Fernando Luís Almeida has a PhD in Computer Science Engineering from Faculty of Engineering of University of Porto (FEUP). He holds also MSc in Innovation and Entrepreneurship and MSc in Informatics Engineering from FEUP. He has around 7 years of teaching experience at higher education levels in the field of computer science and management. He has also worked for 10 years in several positions as software engineer and project manager for large organizations and research centers like Critical Software, CICA/SEF, INESC TEC and ISR Porto. During that time he had the possibility to work in partnership with big international organizations and universities in several European projects. His current research areas include innovation policies, entrepreneurship, software development and decision support systems.

Mohammad Nabil Almunawar is currently a senior assistant professor at School of Business and Economics, Universiti of Brunei Darussalam (UBDSBE), Brunei Darussalam. He was the former dean of UBDSBE. He received his bachelor degree in 1983 from Bogor Agricultural University, Indonesia, master Degree (MSc Computer Science) from the Department of Computer Science, University of Western Ontario, London, Canada in 1991 and Ph.D. from the University of New South Wales (School of

Computer Science and Engineering, UNSW) in 1998. Dr Almunawar has published more than 80 papers in refereed journals, book chapters and international conferences. He has more than 30 years teaching experiences in the area computer and information systems. His overall research interests include applications of IT in Management, Electronic Business/Commerce, Health Informatics, information security, cloud computing, Big Data and Internet of Things. He is also interested in object-oriented technology, databases and multimedia retrieval.

Justinah Amoko is Lecturer at Department of Chemistry, Adeyemi College of Education Ondo.

Maria Teresa Artese took her degree in Computer Science at the University of Studies of Milan in 1990. She has been working at the CNR since 2000. Now she works at IMATI – CNR Unit of Milan. Dr. Artese major areas of work are functional analysis and software development, technical support, database structuring and development, dynamic web database sites, design, and implementation. Recently she has focused his research on multimedia information systems development and data online management. On these topics, she has several national and international publications, and she has been working in national and international research projects.

Artem Artyukhov is a member of Processes and Equipment of Chemical and Petroleum-Refineries Department. Research areas: 1. Mathematical modeling of the main processes and devices of chemical technology. 2. Optimization calculations of equipment for chemical production. 3. Creation of three-dimensional models of technological units and equipment. 4. Computer simulation of hydromechanical and heat-mass transfer processes in chemical technology. 5. Development of new types of heat-mass transfer equipment with the use of swirling flows.

Nadiia Artyukhova is a member of Processes and Equipment of Chemical and Petroleum-Refineries Department. Research areas: 1. Mathematical modeling of the main processes and devices of chemical technology. 2. Optimization calculations of equipment for chemical production. 3. Computer simulation of hydromechanical and heat-mass transfer processes in chemical technology. 4. Development of technological and design documentation for the manufacturing of equipment for chemical production. 5. Development of new types of drying equipment.

Udo Richard Franz Averweg is employed as an Information Technology (IT) Project Manager at eThekwin Municipality, Durban, South Africa. He entered the IT industry during 1979 and holds a Masters Technology degree in Information Technology (cum laude), a second Masters degree in Science from the University of Natal and a third Masters degree in Commerce from the University of KwaZulu-Natal, Durban, South Africa. As an IT practitioner, he is a registered professional member of the Institute of Information Technology Professionals in South Africa. He has authored and co-authored more than 190 research outputs (90 being peer-reviewed): some research outputs have been delivered at local and international conferences, some have been published in accredited peer-reviewed journals and some have appeared as chapters in books. During January 2000 Udo climbed to the summit of Africa's highest peak, Mount Kilimanjaro (5,895 metres), in Tanzania.

Samuel Oyelami Babalola works in Sales and Marketing department of Livestock Feeds Plc (subsidiary of UAC Nigeria Plc), Lagos, Nigeria. He has previously functioned as a Research & extension Officer,

About the Contributors

and Data analyst in different organisations. He holds a Bachelor of Agriculture in Agricultural Extension and Rural Development from the University of Ibadan, Nigeria. He also holds a Master in Information Science degree (M.Inf.Sc) from the University of Ibadan, Nigeria. His research interest includes Business development and Market research, ICT and society, Business informatics and Knowledge Management.

Grace Babcock received her B.A. in Sociology from North Carolina State University in May of 2019. She will receive her J.D. from the University of South Carolina School of Law in May of 2022. Her research interests include maternal employment and gender disparities in the legal profession.

Abira Banik completed her B.E. degree in Computer Science & Engineering from Tripura Institute of Technology, Agartala in 2015. She has completed M.Tech. in Computer Science & Engineering from Department of Computer Science & Engineering, Tripura University (A Central University), Suryamaninagar, India in 2018. Her research interest includes Wireless Mesh Network, MANET and QoS provisioning.

Diane Barrett is an associate professor. She has done forensic, incident response, and security assessment work for many years holding titles such as Manager of research and training for cyber, mobile forensic training director, and enterprise engineer. She has served as a program director for ADFSL's Conference on Digital Forensics, Security, and Law and is a volunteer on the National CyberWatch Center's Curriculum Standards Panel as well as NIST's Cloud Computing Forensic Science Working Group. Dr. Barrett served for several years as the President of the Digital Forensics Certification Board. Dr. Barrett's education includes a Ph.D. in business administration with a specialization in information security. She holds many industry security certifications and has authored multiple security and forensic books.

Amit Bhagat is working as an Assistant Professor at the National Institute of Science & Technology, Bhopal. His expert area is Big Data, Machine Learning, and IoT. More than 10 papers are published in international journals and presented many papers at international conferences. He guided many research scholars who are working in various areas of computer science.

Tulshi Bezboruah (M'12) received the B.Sc. degree in physics with electronics from the University of Dibrugarh, Dibrugarh, India, in 1990, and the M.Sc. and Ph.D. degrees in electronics and radio physics from the University of Gauhati, Guwahati, India, in 1993 and 1999, respectively. In 2000, he joined in the Department of Electronics Science, Gauhati University, as a Lecturer. He is currently the Professor & Head, Department of Electronics and Communication Technology, Gauhati University. His current research interests include instrumentation and control, distributed computing, and computer networks. Prof. Bezboruah is a senior member of the IEEE, and Vice Chair of IEEE Geoscience and Remote Sensing Society (IEEE GRSS), Kolkata chapter as well as an Associate Member of the International Center for Theoretical Physics, Trieste, Italy.

Cristina Boido has a degree in Architecture, specialization in History, Analysis and Development of Architectural and Landscape Heritage at the Politecnico di Torino, research Doctorate in Drawing and Survey of Architectural Heritage at the Università "La Sapienza" di Roma, associate professor of Drawing at the Department of Architecture and Design of the Politecnico di Torino from December 2010. Currently she teaches a class of Representation in the Territorial, Urban, Environmental and Landscape Planning

Laurea and in the Atelier Compatibility and sustainability of architectural restoration in Master's degree programmes of the Architecture for Sustainable Design. Cristina Boido has in depth knowledge of the problems linked to Drawing and Representation, and in particular to architectural, urban and landscape surveying, using digital representation techniques.

Abhijit Bora received Master of Computer Applications (MCA) degree from Jorhat Engineering College (Under Dibrugarh University), India in 2008 and Ph.D from Gauhati University, India in the year 2017. His research interests include web service, web security and software engineering.

Roula Borhani obtained her Bachelor degree in electronic marketing and digital strategy with honors from Higher School of Digital Economy of Manouba (Tunisia) in 2019. She is preparing a master's degree in management of social and solidarity organizations at the High Institute of Management of Tunis. She practices in the areas of digital marketing and social media. She has more than 10 years of experience as a communication manager and content editor in international companies. She is a member of various non-governmental organizations working in the field of human resource development (OMEDRH, AFK).

Djelloul Bouchiha received his Engineer degree in computer science from Sidi Bel Abbes University, Algeria, in 2002, and M. Sc. in computer science from Sidi Bel Abbes University, Algeria, in 2005, and Ph. D. in 2011. Between 2005 and 2010, he joined the Department of Computer Science, Saida University, Algeria, as a Lecturer. He became an Associate Professor since September 2013. Currently, he is a Full Professor at the University Center of Naama. His research interests include software engineering, semantic Web services, Web reverse-engineering, ontology engineering, knowledge management and information systems.

Charalampos Bratsas teaches at the Mathematics Department of the Aristotle University of Thessaloniki. He is the founder of the Greek Chapter of the Open Knowledge Foundation and the President of its Administration Board. He is a specialist in Semantic Web with a focus on Linked Open Data Internationalization and Ontology engineering. His research interests include Data Mining algorithms, Semantic Web, Social Semantic Web, Internationalization of Linked Data, he is coordinator of Greek DBPedia, Applications of Statistical Knowledge, Data Journalism, as well as Neuroinformatics. His research has appeared in the Journal of Web Semantics: Science, Services and Agents on the World Wide Web, in IEEE Transactions on Information Technology in Biomedicine, in the Journal of Computer Methods and Programs in Biomedicine, in BMC Medical Education, and others.

Sheldondra Brown is a student at Southeast Missouri State University to receive her Bachelor of Science in Psychology. Some of her research interest include cultural identity, group dynamics, race relations, cognitive psychology, and mental health outreach.

Tomas Cahlik is Associate Professor of Economics at the Institute of Economic Studies, Charles University in Prague, Czech Republic. His research and teaching activities are currently in Comparative Economics and Ethics and Economics.

Pedro Campos is an assistant professor at the Faculty of Economics of the University of Porto (UP), with a PhD in Management Sciences by U. Porto, Bsc in Applied Mathematics (U. Portucalense). He is

About the Contributors

a member of LIAAD (Laboratory of Artificial Intelligence and Decision Support), which is a research group belonging to INESC TEC. He coordinates the Post-Graduate Programme on Business Intelligence and Analytics at the University of Porto Business School. Pedro is actually Deputy Director of the international Statistical Literacy Project (ISLP) and Vice President of IASE - International Association of Statistical Education His main research interests are Statistics (Data Analysis, Sampling Techniques, Statistical Education) and Artificial Intelligence (Multi-Agent Systems). The main areas of applications are Demography and Organizational Networks.

Manuel Caro is Head of Department of Educational Informatics Universidad de Córdoba Professor in Educational Informatics - Unicordoba Software Engineering Specialist - Udistrital Bogotá Master in ICT for education - IUP - Spain Ph.D. Engineering - Systems and Informatics - UNAL – Medellín.

A. R. Chabukswar completed his B. Pharm. (1998) from Pune University, M. Pharm. (2000) from Mumbai University and Ph. D. (2009) from Shivaji University Kolhapur. He is presently working as Professor & HOD Pharmaceutical Chemistry at School of Pharmacy, Dr. Vishwanath Karad MIT world Peace University, Pune from 19th September 2000. He is actively involved in Academic, research and administrative activities of Pharmacy school. He has published more than 60 research articles in referred Journals. He has co-authored 5 books. He has filed 2 Indian patents. He has completed 2 minor research projects of Savitribai Phule Pune University. He has received research grants from DST and CSIR. He has received Best Professor in Pharmaceutical Chemistry Award in 11th Innovative Education Leadership Award by DNA on 16th February 2019. Awarded as Ideal Teacher award by MAEER's Maharashtra Institute of Technology in 2008. He has guided 47 M.Pharm students and 02 Ph.D students. Presently 06 students of M.Pharm and 03 students of Ph.D. are working on their project under his guidance.

Pankaj Chaudhary is currently Professor of MIS in the ISDS department at Indiana University of Pennsylvania, USA. He holds a PhD in MIS from Southern Illinois University, USA, an MBA from Indian Institute of Management, Ahmedabad, India, and Bachelor of Technology in Computer Science & Engineering from Indian Institute of Technology, Delhi, India. His current research areas include Information Systems Agility, IS Infrastructure, IS-Centric organizational transformation, and Knowledge Management.

Nutan Kumari Chauhan is an Assistant Professor at KIET Group of Institution, Ghaziabad, India. She did her Ph.D. in 2019 from Gurukula Kangari Vishwavidyalaya Haridwar, 249404, Uttarakhand India. Her area of research is Distributed Fuzzy System. She has also qualified CSIR NET-JRF. She has published five research paper in International Journals.

E. Jack Chen is a Senior Staff Specialist with BASF Corporation. He received a Ph.D. degree from University of Cincinnati. His research interests are in the area of computer simulation.

Dilip Kumar Choubey received his Ph.D in Engineering from Birla Institute of Technology (B.I.T), Mesra, Ranchi, India in 2018. He received his M.Tech degree in Computer Science and Engineering from Oriental College of Technology (O.C.T), Bhopal, India in 2012 and has B.E. degree in Information Technology from Bansal Institute of Science and Technology (B.I.S.T), Bhopal, India in 2010. Currently, He is working as an Assistant Professor (On Contract) in NIT Patna, India since March 2017. He

worked as an Asst. Prof. in Lakshmi Narain College of Technology (L.N.C.T), Bhopal, India and Oriental College of Technology (O.C.T), Bhopal, India. He has more than 7 years of teaching and research experience. He is a author of 1 book and has more than 19 research publications in international journals, book chapters and conference proceedings to his credit. He has been the member of the organizing and technical program committees of some conferences and Workshop. He has reviewed several research papers in reputed journal. He has been attended so many workshop's/Seminar's/Conference's. He is the lifetime members of several professional bodies such as The International Association of Engineers (IAENG), The Internet Society Switzerland (ISOC), The Society of Digital Information and Wireless Communications (SDIWC), Universal Association of Computers and Electronics Engineers (UACEE), Associate Member (till 4 May, 2019), Computer Science Teachers Association (CSTA), etc. His research interests include Machine Learning, Soft Computing, Bioinformatics, Data Mining, Pattern Recognition, and Database Management System, etc.

Mahfuzulhoq Chowdhury received the BSc degree in computer science and engineering from Chittagong University of Engineering and Technology, Bangladesh, in 2010, and the MSc degree in computer science and engineering from Chittagong University of Engineering and Technology, in January 2015. He is currently working toward the PhD degree in INRS working on the Tactile Internet. His current research interests include collaborative computing, fiber-wireless multi-robot networks, human-to-robot communications, and Tactile Internet.

Nkiko Christopher is the current University Librarian of Elizade University. Prior to joining the services of Elizade University, he had a stint for a year at Samuel Adegboyega University in the same capacity after a successful tenure as the University Librarian of Covenant University. He holds B.Sc in Political Science from the University of Benin, Edo State, M.Sc in Political Science, MLS and Ph.D in Library and Information Studies from the University of Ibadan. He holds Chief Oduaran Memorial Prize (1988) for being the best graduating student in Political Science, Chief Anthony Enahoro's Prize for being the best overall student in Nigerian Government and Politics (1988) and NYSC, Kano State Merit Award for Exemplary Character and Distinguished Hardwork (1989), LARISSA, University of Ibadan (2012) Distinguished Personality Award for Selfless Service to Librarianship and Information Profession, Award for Effective Archival System and Valid Contribution to Scholarship in Covenant University (Covenant University 2014-2015 Awards and Recognition). Dr. Nkiko has had the privilege of international study visits to benchmark the following United Kingdom and American libraries: British Library (London), Lancaster University Library, Birmingham City University Library, Aston University Library, Birmingham, University of Sussex, Brighton, Fayetteville State University Library, North Carolina State University Library, North Carolina Central University Library and Calvin College Library, Grand Rapids, Michigan. He is an Associate member of the Chartered Institute of Library and Information Professionals (CILIP) UK, member of the Nigerian Library Association, member of the Association of University Librarians of Nigerian Universities and a Certified Librarian (CLN) of the Librarians' Registration Council of Nigeria. Dr. Nkiko has published extensively in local and international (SCOPUS and Web of Science-indexed) journals. He is on the editorial boards of many reputable journals.

Vera Costa has a degree in Mathematics at the Faculty of Sciences of the University of Porto and a Master Degree in Data Analysis and Decision Support Systems at the Faculty of Economics of the same university. She started to teach mathematics to high school and, after that, information systems to

About the Contributors

college students. She participated in several research projects, and she contributes with her knowledge of statistical analysis and software programming, in different application fields, such as health, politics, and transportation. Currently, she is a Ph.D. student of Transportation Systems of the MIT Portugal program.

Kevin Curran is a Professor of Cyber Security, Executive Co-Director of the Legal innovation Centre and group leader of the Ambient Intelligence & Virtual Worlds Research Group at Ulster University. His achievements include winning and managing UK & European Framework projects and Technology Transfer Schemes. Professor Curran has made significant contributions to advancing the knowledge and understanding of computer networks and security, evidenced by over 800 published works. His expertise has been acknowledged by invitations to present his work at international conferences, overseas universities and research laboratories. He is a regular contributor to print, online, radio & TV news on computing & security issues. He was the recipient of an Engineering and Technology Board Visiting Lectureship for Exceptional Engineers and is an IEEE Public Visibility technical expert since 2008. He currently holds a Royal Academy of Engineering/Leverhulme Trust Senior Research Fellowship awarded in 2016. Professor Curran's stature and authority in the international community is demonstrated by his influence, particularly in relation to the direction of research in computer science. He has chaired sessions and participated in the organising committees for many highly-respected international conferences and workshops. He was the founding Editor in Chief of the International Journal of Ambient Computing and Intelligence and is also a member of numerous Journal Editorial boards and international conference organising committees. He has authored a number of books and is the recipient of various patents. He has served as an advisor to the British Computer Society in regard to the computer industry standards and is a member of BCS and IEEE Technology Specialist Groups and various other professional bodies.

Arpita Das received her B. Tech, M.tech degrees in Radio Physics and Electronics University of Calcutta, India in the year 2004 and 2006 respectively. She received her Ph.D. degree in 2012 from University of Calcutta. Currently she is an Assistant Professor in the Dept. of Radio Physics and Electronics.

Kaustav Das is an M.Tech student in the Dept. of Radio Physics and Electronics, University of Calcutta, India. He received his B.Tech degree in Radio Physics and Electronics in the year 2018.

Aristides Dasso is currently a Visiting Professor at the Departamento de Informática at Universidad Nacional de San Luis, Argentina. He has degrees from Universidad Nacional del Litoral, Rosario, Argentina (Political Science); Université de Paris VII, Paris, France (Computer Science) and Universidad Nacional de San Luis (Master in Software Engineering). He has been a fellow at the International Institute for Software Technologies of the United Nations University, Macau; and was at Universitat Politècnica de València, Spain, with a scholarship from the European Community ALFA LERnet program. He is a member of ACM, IEEE, and SADIO of Argentina. He have published more than fifty articles in international conferences and journals on Software Engineering and Information Systems, areas of research that he is currently pursuing, with special emphasis in the area of Formal Methods and Machine Learning. He have served and is serving as referee in several conferences both national and international.

Pia Davico is Architect, Adjunct Professor of Drawing and Representation, and Assistant Professor at the Politecnico di Torino, Department of Architecture and Design (DAD); Ph.D. in "Drawing and survey of Architectural Heritage" at the Università "La Sapienza" di Roma. Teacher in the "School of

Specialization on Architecture and Landscape Heritage” of Politecnico di Torino, in Ateliers of the master degree course “Architecture for the sustainable projects” and in the “Laboratory of Drawing and Survey”. Reviewer for some important international magazines. She participated in scientific research groups, in consulting projects for public Institutions, in National and International Conferences with numerous oral relations and written researches. She always privileged the interest for architecture survey inserted in its own context, which provided a variety of original experiences through collaborations with the Public Administrations. The survey considered the study of the historical transformations of the built and the on-site analysis of urban structure to different scales: from relationship with the local territory to buildings. In this way the environmental and architectural characters were identified, including the color, with perceptive and functional studies.

José C. Delgado is an Associate Professor at the Computer Science and Engineering Department of the Instituto Superior Técnico (University of Lisbon), in Lisbon, Portugal, where he earned the Ph.D. degree in 1988. He lectures courses in the areas of Computer Architecture, Information Technology and Service Engineering. He has performed several management roles in his faculty, namely Director of the Taguspark campus, near Lisbon, and Coordinator of the B.Sc. and M.Sc. in Computer Science and Engineering at that campus. He has been the coordinator of and researcher in several international research projects. As an author, his publications include one book, 25 book chapters and more than 50 papers in international refereed conferences and journals.

Suparna Dhar is a senior manager in an IT organization with focus on developing IT solutions for the Fintech space and exploring emerging technologies. She takes interest in academic research on application, adoption and use of emerging technologies in business, covering analytics, social media technologies, cloud computing, distributed processing and blockchain. She is pursuing her PHD on Information Systems Management from Aligarh Muslim University.

Ljubica Diković is an Assistant Professor at the Faculty of Applied Economics, Management and Finance at the University of Business Academy Novi Sad, and a Professor of vocational studies at the College of Applied Studies Uzice. She was awarded MSc degree by the Faculty of Science, University of Belgrade. She enrolled in PhD studies at the Faculty of Science, University of Kragujevac and acquired the title of PhD in Mathematical Sciences. She has participated in a number of projects funded by the European Union. She has been the author of several articles published in international and national journals, and of more than 40 papers written for international and national conferences.

Gregor Donaj made his B.Sc. degrees in Electrical Engineering and Mathematics in 2010 and 2011 respectively. He obtained his Ph.D. in 2015. From 2010 to 2014 he was a researcher, and from 2014 onward he is a teaching assistant at the Faculty of Electrical Engineering and Computer Science of the University of Maribor. His primary research areas are language modeling and error analysis in speech recognition and machine translation.

Jorge Doorn is Full Professor of Universidad Nacional del Oeste and Universidad Nacional de Tres de Febrero, Argentina. His current research interest is on Requirements Engineering field.

About the Contributors

Dinah Duarte is Professor at Universidade Lusófona's Research Center for Biosciences and Health Technologies, Portugal.

Robert Andrew Dunn is assistant chair for the Department of Media and Communication at East Tennessee State University. He is also an associate professor and head of the journalism program. He teaches both journalism and mass communication courses. Academically, Dunn holds a Ph.D. in mass communication from the University of Alabama, a master's degree in mass communication from Florida State University, and a bachelor's degree in mass communication from Auburn University Montgomery. Professionally, Dunn worked in newspapers for nearly a decade. He served primarily as an education reporter and occasional legislative reporter for both the Tallahassee Democrat in Tallahassee, Florida, and The Ledger in Lakeland, Florida. He has also served as an instructor for Florida Southern College and the University of Alabama. Dunn's research focuses on avatars, virtual identity, mediated identity, media psychology, media effects, fandom, and communication technology. He also has a strong interest in journalism and media convergence. He is an avid gamer, dedicated sports fanatic, expert comic collector, enthusiastic sci-fi/fantasy devotee, frequent fan convention attendee, and Belgian beer buff.

Justina Ngozi Ekere a Chartered Librarian, is presently a Principal Librarian and Head of Reference Section, Nnamdi Azikiwe Library University of Nigeria Nsukka. She is a holder of B.Sc. (Hons); M.L.S and a Ph.D in Library and Information Science. She is a multifaceted library practitioner, educator and also a seasoned administrator. In 1997, she was employed as an Assistant Librarian and today she has over Twenty-Three (23) years post qualification and cognate experience. Dr. Ekere is creative and thorough. She has initiative and a cheerful disposition which enable her to get on with her superiors and peers. In 2004, the Governing Board of the American Biographical Institute (ABI) in recognition of her significant accomplishments within, and mastery of library and Information Science, conferred on her GREAT WOMAN OF THE 21ST CENTURY, a title reserved for women whose accomplishments and influence are the result of superior dedication and professional integrity. She is a member of professional associations such as Nigerian Library Association (NLA) AFLIA, IFLA, African Centre of Excellence for Information Ethics (ANIE) and WALA). Her Research and Scholarly Interest includes; Indigenous Knowledge Management, Human Resource Management, Information Ethics and ICT applications to library operations.

Abeer Elshater PhD, is a Professor of Urban Morphology at Ain Shams University (ASU), Cairo, Egypt, teaching and supervising multidisciplinary topics in the ideology of urban design. Her area of interest is in urban design disciplines. She has worked on some international research projects with international universities. Elshater is an author and co-author of seven books as well as a number of scientific manuscripts in international journals and conference proceedings. Recently, she acts as an ambassador of the Regional Studies Association, United Kingdom.

Arzu Erener graduated from the Department of Geodesy and Photogrammetry Engineering at Istanbul Technical University (ITU) in 1999, she has completed a master's education while working as a research assistant in the Geodesy and Photogrammetry Department at Karadeniz Technical University (KTU) between 2000-2002. Researcher completed her doctorate from the Middle East Technical University (METU), in 2010, while serving as a research assistant between 2002-2009. She worked as research assistant doctor in Selcuk University (SU) Geomatics Engineering Department and got her professorship

at 18 January 2013. Currently she is working as Assoc. Prof. Dr. at Geomatics Engineering Department of Kocaeli University and gives courses on Remote Sensing, GIS and Risk Management. Since 2015, she has been working as the head of Geomatic Engineering Department.

Monday Ofori Eyinagho holds a B.Sc. Degree in Electronic and Electrical Engineering, an M.Sc. Degree in Computer Engineering, an MBA Degree in General Management, and a Ph. D in Computer Engineering. He is currently a Senior Lecturer in the Department of Electrical and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria. His principal researches are in the areas of development and implementation of microprocessor-based systems, performance optimization of Local Area Computer Networks, Wireless Sensor Networks, Internet of Things, Digital and Microprocessor-based Instrumentation Systems. He is a Life Member of the Nigeria Computer Society, Member of the Nigerian Society of Engineers, a Council for the Regulation of Engineering in Nigeria Registered Electronic Engineer.

Michael O. Fagbohun is a seasoned professional librarian with an unusual passion for youth mentoring. He is a multi-talented individual that seamlessly combines professionalism with research and social entrepreneurship. He holds a Bachelors Degree in Library and Information Science (B.Sc Edu.) from Tai Solarin University of Education, Ogun State and Masters Degree in Library and Information Science (MLIS) from the prestigious University of Ibadan, Nigeria. He is a member of the Nigerian Library Association and a certified librarian of the Library Registration Council of Nigeria. In recognition of his drive for excellence and quest for youth mentoring, Michael won the Sokoto State Honours Award as Best Volunteer for the Millennium Development Goals during his service year and best ENACTUS Mentor. His flair for social entrepreneurship has propelled him into personally executing a number of community impact projects and also coordinated and administered a wide range of same under the auspices of ENACTUS (as Faculty Advisor). These projects have earned his team several awards. His involvement with ENACTUS for over ten years made it possible for him to introduce same to Covenant University in 2016. He is also a volunteer trainer for Microsoft and African Development Bank project – coding for employment program in Nigeria. He has personally founded MeentuP (online mentorship platform for academia, teachers and entrepreneurs), and People's Varsity - a community social change hub that offers individuals in disadvantaged areas access to learning space and information resources. He is presently working on how to make research output of Africans more visible and developed a project is tagged IBILEDATA. Professionally, Michael has pioneered two academic libraries - Iresi College of Technology, Osun State in 2013 and St. Albert The Great Major Seminary, Idow-Offoron, Abeokuta, Ogun State in 2015. Currently, he works at Covenant University as a Librarian/Faculty at the Centre for Learning Resources. As an academic, Michael is well published and has to his credit over 25 publications in highly reputable international journals, conference proceedings as well as book chapters. His research interest revolves around distributed digital libraries, metadata, Knowledge Organization Systems (KOS), Semantic Web technologies, institutional repository, open source software and community and lifelong learning.

Samuel Oluwale Falaki is the current Head, Department of Electrical and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria. He is a Fellow, Nigeria Computer Society, and Computer Professionals Registration Council of Nigeria; Member, Nigeria Society of Engineers, and he is an Engineer registered with the Council for the Regulation of Engineering in Nigeria.

About the Contributors

Liliana Favre is a full professor of Computer Science at Universidad Nacional del Centro de la Provincia de Buenos Aires in Argentina. She is also a researcher of CIC (Comisión de Investigaciones Científicas de la Provincia de Buenos Aires). Her current research interests are focused on model driven development, model driven architecture and formal approaches, mainly on the integration of algebraic techniques with MDA-based processes. She has been involved in several national research projects about formal methods and software engineering methodologies. Currently she is research leader of the Software Technology Group at Universidad Nacional del Centro de la Provincia de Buenos Aires. She has published several book chapters, journal articles and conference papers. She has acted as editor of the book UML and the Unified Process. She is the author of the book Model Driven Architecture for Reverse Engineering Technologies: Strategic Directions and System Evolution.

Ourania-Ioanna Fotopoulou is a family physician/general practitioner. She holds a MD from L' Aquilla University, Italy and a MSc in Primary Health Care from the Aristotle University of Thessaloniki, Greece.

Alexander Fridman graduated from Leningrad Electro-technical Institute in 1975 and worked in Baku (Azerbaijan) for Russian Ship-building Ministry until 1989, when he moved to Apatity (Murmansk region, Russia) and began working for Russian Academy of Sciences (RAS). He got his PhD in 1976, Doctor of Science (Technical) degree in 2001 and Professor degree in 2008. At present he is a Leading Scientific Researcher in the Institute for Informatics and Mathematical Modelling (IIMM) of the Kola Science Centre of RAS and a Professor of the Information Systems Chair in the Kola Branch of the Petrozavodsk State University. Alexander Fridman is a member of the Russian Association for Artificial Intelligence and an expert of the Russian Foundation for Basic Researches. He has 270 scientific publications including 5 monographs, 24 tutorials and 16 certificates for inventions. His main scientific interests are in logical intelligence systems and modelling of hierarchical systems.

Ana Funes is a Professor with tenure at the Departamento de Informática of the Universidad Nacional de San Luis, San Luis, Argentina. She has got a Master degree in Software Engineering and Formal Methods from the Universidad Politécnica de Valencia, Valencia, Spain. She has also been a fellow at the International Institute for Software Technologies of the United Nations University, Macau, where she worked on the integration of formal and semi-formal notations. Her research interests include the combined use of formal and traditional methods for software system specification. She is also interested in Requirement Engineering and Machine Learning. She has published more than fifty articles on these topics in international conference proceedings, books and international journals.

Isabella Gagliardi works at IMATI- CNR since September 2016. She has been working at CNR since 1986. Her major areas of research include Hypermedia Information Retrieval models and methodologies, automatic generation of hypertextual links between text/audio/image, dynamic web based database design and implementation, and clustering algorithms. Recently he has focused his research on the development of multimedia information systems available on the web. On these topics she has several national and international publications, and she has been working in national and international research projects.

Stéphanie Gauttier is an Assistant Professor at Grenoble Ecole de Management. She received her PhD in Information Systems from the University of Nantes (France). She conducts research at the cross-roads between ethics and information systems. Her fields of expertise concern human enhancement, augmented reality, responsible innovation, and Q-methodology.

Beata Gavurova is an expert in finance, financial analysis and financial risk management. She focuses in her research work on issues of measurement and performance management in various sectors, process management and process optimization, strategic and performance benchmarking. Her dominant research area is the development and testing in the management and performance measurement, the evaluation of methodologies and the preparation of an application platform for innovative management and performance measurement with the support of ICT. She led several national projects, has participated in many international projects targeting the innovative applications of ICT in the private and public sector.

Alexander Gegov is currently Reader in Computational Intelligence in the School of Computing. He has been Associate Dean Research for the Faculty of Technology in the past. He has a PhD in Control Systems and a DSc in Intelligent Systems – both from the Bulgarian Academy of Sciences. He has been a recipient of a national award for the best young researcher from the Bulgarian Union of Scientists. He has been Humboldt Guest Researcher at the Universities of Duisburg and Wuppertal in Germany as well as EU Visiting Researcher at the Delft University of Technology in the Netherlands.

James A. George is an experienced CEO, a former executive for Fortune 500 Companies, a consultant for PriceWaterhouseCoopers and ManTech International. He consulted to the US Department of Defense and Health and Human Services in information technology. He served as Academic Director for UC Berkeley Worldwide, developing and instructing course in business management for Asia-nation students. George published hundreds of articles and papers including contributing the Asia Pacific Economic Cooperative Conference on Information Literacy. George was Vice President and Publisher for the Hitchcock Division of the American Broadcasting Companies and Publisher of Infosystems Magazine. George is a graduate of Franklin University and the Ohio State University. Today, George is an independent journalist and author having contributed to CNN and CBS News. Co-Author: Smart Data, Enterprise Performance Automation Strategy © 2010 Wiley Publishing; How to Select an American President © 2017 Archway Publishing; Regenerating America with Sustainable Economics © 2017 Archway Publishing. Author: A President's Manifesto © 2019 All Rights Reserved.

Vyshali Gogi is a VTU Research Scholar at RV College of Engineering®, Bengaluru, Karnataka, India. She holds BSc degree in Computer Application and post graduate degree in "Master of Computer Applications (MCA)" with distinction from University of Mysore, Karnataka, India. Her research area is "Data Mining in Healthcare".

Adán Gómez is Professor of Educational Informatics Department of Educational Informatics (Departamento de Informática Educativa) Universidad de Córdoba.

Patrícia Gonçalves is a graduate in Applied Physics/Mathematics (Astronomy) and master in Computational Methods in Sciences and Engineering. Part of INESC-TEC since October 2016, developing

About the Contributors

work in Network Science, more precisely in multilayer networks. Student of the PhD Program in Engineering and Public Policy.

Sotirios K. Goudos received the B.Sc. degree in Physics in 1991 and the M.Sc. of Postgraduate Studies in Electronics in 1994 both from the Aristotle University of Thessaloniki. In 2001, he received the Ph.D. degree in Physics from the Aristotle University of Thessaloniki and in 2005 the Master in Information Systems from the University of Macedonia, Greece. In 2011, he obtained the Diploma degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki. From 2004 to 2007, he was a part-time lecturer with the Department of Technology Management, University of Macedonia, Greece. Since 2013, he has been with the Department of Physics, Aristotle University of Thessaloniki. His research interests include antenna and microwave structures design, electromagnetic compatibility of communication systems, evolutionary computation algorithms, and mobile communications. Dr. Goudos is a Senior member of the IEEE, the Greek Physics Society, the Technical Chamber of Greece, and the Greek Computer Society.

M. Govindarajan is currently an Assistant Professor in the Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India. He received the B.E, M.E and Ph.D Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2001, 2005 and 2010 respectively. He did his post-doctoral research in the Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, United Kingdom in 2011 and at CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore in 2013. He is pursuing Doctor of Science Degree in Computer Science and Engineering at Utkal University, Bhubaneswar, Orissa, India Since 2013. He has visited countries like Czech Republic, Austria, Thailand, United Kingdom (twice), Malaysia, U.S.A (twice), and Singapore. He has presented and published more than 100 papers at Conferences and Journals and also received best paper awards. He has delivered invited talks at various national and international conferences. His current research interests include Data Mining and its applications, Web Mining, Text Mining, and Sentiment Mining. He has completed two major projects as principal investigator and has produced three Ph.Ds. He was the recipient of the Achievement Award for the field in the Conference in Bio-Engineering, Computer Science, Knowledge Mining (2006), Prague, Czech Republic. He received Career Award for Young Teachers (2006), All India Council for Technical Education, New Delhi, India and Young Scientist International Travel Award (2012), Department of Science and Technology, Government of India, New Delhi. He is a Young Scientists awardee under Fast Track Scheme (2013), Department of Science and Technology, Government of India, New Delhi and also granted Young Scientist Fellowship (2013), Tamil Nadu State Council for Science and Technology, Government of Tamil Nadu, Chennai. He also received the Senior Scientist International Travel Award (2016), Department of Science and Technology, Government of India. He has published two book chapters and also applied patent in the area of data mining. He is an active Member of various professional bodies and Editorial Board Member of various conferences and journals. He is IQAC Coordinator and Nodal Officer for data collection, Department of Computer Science and Engineering and also Member of Patent Cell, Annamalai University.

Graciela Hadad has a degree in Information Systems Bachelor at Universidad de Buenos Aires (Argentina, 1983), and has received a Ph.D. in Computer Science at Universidad Nacional de La Plata (Argentina, 2008). She has conducted software developer teams in several Argentine companies during

the past thirty years. She has been teaching Software Engineering in under-graduate courses for the last thirty-two years at different universities. She has been participating in research projects since 1995. Currently, she is a full professor and researcher at Universidad Nacional del Oeste, while also teaching at Universidad Nacional de La Matanza and Universidad de Belgrano. Since 2003, she has been teaching Requirements Engineering in post-graduate courses at different universities. She is supervisor and co-supervisor of Master Thesis and Doctoral Thesis. She currently belongs to a Requirements Engineering research group at Universidad Nacional del Oeste and at Universidad de Belgrano. She has published several papers in conference proceedings, journals, and book chapters along these years.

Osman Hasan received the B.Eng. (Hons.) degree from the N.W.F.P University of Engineering and Technology, Pakistan, in 1997, and the M.Eng. and Ph.D. degrees from Concordia University, Montreal, Canada, in 2001 and 2008, respectively. He served as an ASIC design Engineer from 2001 to 2003 at LSI Logic Corporation in Ottawa, Canada and as a Research Associate at Concordia University, Montreal, Canada, for 18 months after his doctoral degree. Currently, he is an Associate Professor and Principal and Dean at the School of Electrical Engineering and Computer Science, National University of Science and Technology (NUST), Islamabad, Pakistan. He is the founder and director of System Analysis and Verification (SAVe) Lab at NUST, which mainly focuses on the design and formal verification of embedded systems. He has received several awards and distinctions, including the Pakistan's Higher Education Commissions Best University Teacher (2010) and Best Young Researcher Award (2011) and the Presidents gold medal for the best teacher of the University from NUST in 2015. Dr. Hasan is a Senior member of IEEE, member of Association for Automated Reasoning (AAR) and member of the Pakistan Engineering Council (PEC).

Craig Hight is an Associate Professor in Creative Industries at the University of Newcastle, NSW. His research has drawn on documentary theory, software studies, critical data studies and a variety of approaches within the field of audience research. His most recent work explores the nature of documentary culture and practice within digital media platforms. More details on his website: <https://www.newcastle.edu.au/profile/craig-hight>.

Barbara Holland is a native New Yorker with an ongoing quest for knowledge and information from a variety of disciplines. Holds a Bachelor of Arts Degree in Psychology from the City College of New York, a Master's degree in Library Science from Queens College of New York. She currently works at the Brooklyn Public Library as a Senior Librarian and has over 20 years of experience as a librarian and Information Professional. Presented three years at the Georgia International Conference on Information Literacy and The Brooklyn Reading Council each. Numerous Book chapters published with IGI-Global and the Editor of Emerging Trends and The Impacts of Internet of Things and Libraries. Current Editor of Forthcoming book "Libraries During a Global Health Crisis: Response to the COVID-19 Pandemic." As a librarian and Information Professional, it is imperative to keep up with library research and technological trends and provide the information on an ongoing basis.

Wen-Chen Hu received a BE, an ME, an MS, and a PhD, all in Computer Science, from Tamkang University, Taiwan, the National Central University, Taiwan, the University of Iowa, Iowa City, and the University of Florida, Gainesville, in 1984, 1986, 1993, and 1998, respectively. He is currently an associate professor in the School of Electrical Engineering and Computer Science of the University of

About the Contributors

North Dakota, Grand Forks. He was an assistant professor in the Department of Computer Science and Software Engineering at the Auburn University, Alabama, for years. He is the general chairs of about 20 international conferences such as the 2018 International Conference on Engineering, Technology, and Applied Science (ICETA 2018), and associate editors of several journals like Journal of Information Technology Research (JITR). In addition, he has acted more than 100 positions as editors and editorial advisory/review board members of international journals/books, and track/session chairs and program committee members of international conferences. He has also won a couple of awards of best papers, best reviewers, and community services. Dr. Hu has been teaching for more than 20 years at the US universities and over 10 different computer/IT-related courses, and advising/consulting more than 100 graduate students. He has published over 100 articles in refereed journals, conference proceedings, books, and encyclopedias, edited more than 10 books and conference proceedings, was the editor-in-chief of the International Journal of Handheld Computing Research (IJHCR) from 2010 to 2017, and solely authored a book entitled “Internet-enabled handheld devices, computing, and programming: mobile commerce and personal data applications.” He is a member of ACM (Association of the Computing Machinery). His current research interests include handheld/mobile/smartphone/spatial computing, location-based services, web-enabled information system such as search engines and web mining, electronic and mobile commerce systems, and web technologies.

Adamkolo Mohammed Ibrahim is currently pursuing a PhD in mass communication programme at Bayero University, Kano, Nigeria. He received his master’s degree in Development Communication at Universiti Putra Malaysia (UPM) in 2017. In 2007, he had his first degree (BA Mass Communication) at the Department of Mass Communication, University of Maiduguri, Nigeria. Currently, he teaches mass communication at the University of Maiduguri. He conducts research and writes in ICT adoption for development, social media, cyberbullying, cyber terrorism/conflict, gender and ICT, gender and conflict and online shopping. He has published several journal articles, book chapters and a few books. His most recent work explores the impacts of fake news and hate speech in Nigerian democracy and proposes a theoretical model as a fact-checking tool. More details on his most recent works and all his other publications can be accessed on his website: <https://unimaid.academia.edu/AdamkoloMohammedIbrahim>. Malam Adamkolo is currently serving as a co-researcher in a research project by The Kukah Centre, Abuja, Nigeria. The proposed title of the research is: “Engaging Local Communities for Peacebuilding, Social Cohesion, Preventing and Countering Violent Extremism in Nigeria’s northeast”. Adamkolo has received Publons Top Reviewer Award in 2018 (for being among the top 1% global peer reviewers in Psychiatry/Psychology). In 2017, Elsevier had awarded him a certificate of outstanding peer review with one of Elsevier’s prestigious journals, Computers in Human Behaviour (CHB) which he reviews for; he also reviews for Emerald’s Journal of Systems and Information Technology (JSIT) and several other journals. Much earlier, from 2000 to 2010, he worked as a broadcast journalist in Yobe Broadcasting Corporation (YBC) Damaturu, and from 2008 to 2010 was deployed to Sahel FM (formerly Pride FM, a subsidiary of YBC Damaturu as DJ-cum-producer/presenter/journalist). From 2008 to 2010, he worked as YBC’s focal person on UNICEF and Partnership for the Revival of Routine Immunisation in Northern Nigeria-Maternal, New-born and Child Health (PRRINN-MNCH). From September to October 2018, he served as a consultant to ManienDanniels (West Africa Ltd.) and MNCH2 programme.

Promise Ifeoma Ilo is the University Librarian, PAMO University of Medical Sciences, Port Harcourt, Nigeria, and the immediate past University Librarian, Covenant University, Ota. Dr. Ilo ob-

tained her Masters and PhD degrees in Library and Information Science from the University of Nigeria Nsukka in (2000) and (2017) respectively. She has worked in both academic and school libraries. Dr Ilo has above forty five publications in reputable local and international journals. She was the Chief Editor, Covenant Journal of Library and Information Science (CJLIS), an online journal of the Centre for Learning Resources, Covenant University Ota. Promise Ilo won the ALCTS Training Award of the American Library Association for librarians (2019). She also holds the awards for the best graduating student (Dept. of Library Science) as well as the best graduating student, School of Education, Federal College of Education Eha-Amufu, (1989). Her research interests are Disaster management in libraries, Information and communication technology in libraries, Health information awareness among rural women, Library user education, among others. Dr Ilo is a member of the Gideons International, Nigeria. She has rendered various services on the platform of the Church Action Committee on HIV/AIDS (CACA) in the Diocese of Lagos West.

Roland Izuagbe works as a Serials Librarian in Center for Learning Resources (The University Library), Covenant University, Ota, Nigeria. Roland holds a Bachelor of Science (BSc) in library and information science from Ambrose Alli University (AAU) Ekpoma, and a master degree in library and information studies (MLIS) from the University of Ibadan, Ibadan. He has published in reputable journals like Library and Information Science Research (LISR), The Journal of Academic Librarianship (JACALIB), International Journal of Disaster Risk Reduction (IJDRR), Library Review (Now, Global Knowledge, Memory and Communication), Journal of Cases on Information Technology (JCIT), among others. His research interests include information and communication technology (ICT), information systems acceptance/adoption, electronic information resources, disaster management in libraries, social media use in libraries and organisational science.

Raheleh Jafari received a B.S. degree in pure mathematics from Islamic Azad University in Shabestar, Iran in 2008 and M.S. degree in Applied Mathematics from Islamic Azad University in Arak, in 2010. She received Ph.D. degree in Automatic Control from National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico, in 2017. In 2017, she worked as a post-doctoral research fellow at the Department of Information and Communication Technology, University of Agder, Grimstad, Norway. Currently, she is working as University Academic Fellow at the School of design, university of Leeds, Leeds, United Kingdom. Her research interests include Artificial intelligence, Fuzzy logic, Robotics and Automation.

Swati Jagdale has completed her B. Pharm. (1998), M. Pharm. (2000) from Pune University and Ph. D. (2009) from Shivaji University Kolhapur. She is presently working as Professor at School of Pharmacy, Dr. Vishwanath Karad World Peace University, Pune. She has published more than 80 research articles in reputed National and International Journals. She has written 5 books and 5 Book chapters. She had filed 3 Indian patents. She had completed 2 minor research projects of Savitribai Phule Pune University (4.9 Lakhs). Presently she is working on research project by Council of Scientific and Industrial Research (CSIR) (Amount 25 Lakh) and AICTE project under MODROB scheme (Amount 17.22 Lakhs). She has guided 61 M.Pharm students and 01 Ph.D. student. Presently 04 students of M. Pharm and 07 students of Ph.D. are working on their project under her guidance. She is also working as member, Board of Studies under Science and Technology for Faculty of Pharmaceutics in Savitribai Phule Pune University from 2019. She has awarded as Award of Excellence in Research in 2019 by DK International Research Foundation, Best Professor In Pharmaceutics Studies by 11th DNA Innovative Education Leadership

About the Contributors

Awards in 2019, Best Teacher Award 2017-18 – Professional Category (Urban) by Savitribai Phule Pune University in 2018, SHIKSHAK SHRI AWARD at National Teacher conference and Presentation by Antarrashtriya Samrasta Manch, Jaipur in 2017, DR. G.R. EKBOTE WOMEN PHARMACIST AWARD of MAHA-APTICON-2016 & PHARMA AWARD-2016 by Association of Pharmaceutical Teachers of India, Maharashtra State in 2016, was nominated and received appreciation trophy for SKB YOUNG PHARMACY TEACHER OF THE YEAR AWARD of MAHA-APTICON-2016 & PHARMA AWARD-2016 by Association of Pharmaceutical Teachers of India, Maharashtra State in 2016, Ideal Teacher award by MAEER's Maharashtra Institute of Technology in 2011. Her research article on Formulation and Evaluation of Gastroretentive Drug Delivery System of Propranolol Hydrochloride was one of the top 5 downloaded articles in AAPS PharmSciTech in 2012. She had also worked as member Board of Studies for Faculty of Pharmaceutics of University of Pune in 2011-15.

Jennifer Joe currently works at the University of Toledo, where she is the Undergraduate Engagement Librarian. Previously, she was the Owensboro Campus Librarian for Western Kentucky University. She received her MLS from Indiana University-Bloomington in 2010, and her MA in Sociology from Western Kentucky University in 2011. Her research interests include community outreach, collection development, information literacy, and the integration of technology into libraries and library practices, with special emphasis on social media. She has previously been published in the Encyclopedia of Information Science and Technology, Fourth Edition (IGI Global 2017) and is co-editing a book entitled Social Media for Communication and Instruction in Academic Libraries, which was published by IGI-Global in March 2019.

Marianna Kapari holds a MSc in Human Computer Interaction from Uppsala University, Sweden, a BSc degree in Computer Science from London Metropolitan University, UK, and a BA in Communication from the American College of Greece. Marianna joined INFOLYSiS as Frontend and digital media developer with experience in web technologies (HTML, CSS, PHP, Javascript, MySQL), prototyping tools (InVision, Balsamiq Mockups, Axure, WireframeSketcher), video editing (Adobe Premier, Ulead VideoStudio, Cyberlink Power Director) and software design tools with UML. Her research interests include IoT interoperability and IoT GW frontends with emphasis on novel HCI technologies, such as Chatbots with Machine Learning techniques.

Shogo Kato is an Associate Professor in the School of Arts and Sciences, Tokyo Woman's Christian University in Japan. He earned a Ph.D. from Tokyo Institute of Technology in 2005. He is particularly interested in the emotional aspects in virtual community, such as Internet bullying.

Yuuki Kato is an Associate Professor at Sagami Women's University, Faculty of Arts and Sciences in Japan. He earned a Ph.D. from Tokyo Institute of Technology in 2005. His general research interests include educational technology, the application of behavior science, psychology, and information and communication technology to educational scenes.

Elaine Khoo is a senior research fellow at the Wilf Malcolm Institute of Educational Research (WMIER) at the University of Waikato, New Zealand. Elaine's research interests include information and communication technology (ICT) and digital pedagogies, e-learning/online learning with a particu-

lar interest in online learning communities, participatory learning cultures and collaborative research contexts. More details on her website: <http://www.waikato.ac.nz/staff-profiles/people/ekhoo>.

George M. Korres holds a doctorate in Economic Sciences (Economics of Innovation) from University of London (QMW) and an M.Sc from Essex University. He is currently a visiting fellow from University of Newcastle, Centre from Urban Regional & Development Studies, (CURDS), School of Geography, Sociology and Politics, and also an associate professor at the University of the Aegean, Department of Geography. He was visiting scholar at the University of Leeds during 2006-2007. Over the last 10 years, during both his academic and professional career, he has gained an in depth experience in the fields of economics, business and politics in the area of European Economics and Innovation Systems. His main areas of interest and specialisation include: European Integration, Social Policy and Social Inclusion, Economic and Innovation Policies, EU enlargement and the Euro-Mediterranean partnership. He has collaborated with many European universities, Institutions, NGOs as well as academia. George Korres has also been involved in various trans-European and international projects and he has published widely in this area in many International Journals, such as *Applied Econometrics*, *Regional and Sectoral Economic Studies*, *Southwestern Economic Review*, *Pensions*, *Sage*, *Management Sciences* and *Regional Development* etc, also he has published a number of books with international publishers, such as MacMillan-Palgrave Press, Springer, Edward and Elgar, Ashgate Publishers and Nova (New York) Publishers.

Maximiliano E. Korstanje is editor in chief of *International Journal of Safety and Security in Tourism* (UP Argentina) and *International Journal of Cyber Warfare and Terrorism* (IGI-GlobalUS). With more than 700 published papers and 25 books, Korstanje was awarded as Outstanding Reviewer 2012. *International Journal of Disaster Resilience in the Built Environment*. University of Salford, UK, Outstanding Reviewer 2013. *Journal of Place Management and Development*. Institute of Place, Manchester Metropolitan University, UK and Reviewer Certificate of Acknowledgement 2014. *International Journal of Contemporary Hospitality Management* (IJCHM), University of Central Florida, US. Now he co-edits almost 10 specialized journals in such themes as human rights, mobility, tourism and terrorism. Korstanje is subject to biographical records for *Marquis Who's Who in the World* since 2009. He had nominated to 5 honorary doctorates for his contribution in the study of the effects of terrorism in tourism. In 2015 he was awarded as Visiting Research Fellow at School of Sociology and Social Policy, University of Leeds, UK and Visiting Professor at University of La Habana Cuba 2016.

Aleksandra Kostić-Ljubisavljević was born 1974. in Belgrade. She received her B.Sc. (1999), M.Sc. (2005) and Ph.D (2011) degrees in telecommunication traffic engineering at the University of Belgrade - Faculty for Transport and Traffic Engineering, where she is currently employed as Associate Professor. She published more than 140 papers regarding her scientific area of interest: economical and technical aspects of telecommunication networks integration, optical communication networks, visible light communication, intelligent transportation systems (communication systems), and telecommunication network regulation.

Harilaos Koumaras was born in Athens, Greece in 1980. He received his BSc degree in Physics in 2002 from the University of Athens, Physics Department, his MSc in Electronic Automation and Information Systems in 2004, being scholar of the non-profit organization Alexander S. Onassis, from the

About the Contributors

University of Athens, Computer Science Department and his PhD in 2007 at Computer Science from the University of Athens, Computer Science Department, having granted the four-year scholarship of National Centre of Scientific Research “Demokritos”. He has received twice the Greek State Foundations (IKY) scholarship during the academic years 2000-01 and 2003-04. He has also granted with honors the classical piano and harmony degrees from the classical music department of Attiko Conservatory. Since 2018, Dr. Koumaras is an Assistant Research Professor at NCSR “Demokritos”.

Vaios Koumaras received his BSc degree in Business Administration with major in Computer Information Systems from the American College of Greece and his MBA in IT Project Management from the City University, USA. Since 1997, he has worked in several positions as Computer Analyst and Software Developer, participating in major IT projects. Currently, he is the managing director of INFOLYSIS P.C., with participation and collaboration in R&D IT projects of numerous companies worldwide, mainly in the development of chatbot apps and the applicability of IoT on top of virtualized infrastructures mainly in the shipping sector for optimizing the cargo containers embarking process and choosing the optimal route, subject to the load and the weather/sea factors. Moreover, Mr. Koumaras is specialized in the development of chatbots over messaging platforms and in the IoT interoperability provision, researching also the applicability with novel SDN/NFV techniques.

Jan Krmela is a head of Department of Numerical Methods and Computational Modelling. Research areas: 1. Computational modeling and tests of composites with a rubber (an elastomer) matrix such as tire casings, especially with focus on strain-stress states and modal analysis and computational modeling of composites after the degradation processes, tire-road interaction and vehicle parts; 2. Tests of tires on static and dynamic test machines with the pressure footprint analyses of contact footprints between tires and a plane road or a bump and also prediction of radial stiffness of tires; 3. Low cyclic loading tests of composites with textiles on a test machine with a video extensometer and static tests of composites with planning of tests with a design of geometric parameters of test samples and test conditions for testing of tire casing parts; 4. Determination of material parameters that can be used as input data into computational models; Microscopic observation of interface bonds between steel and textile cords and an elastomer; 5. 3D printing of technical objects.

Matus Kubak focuses in his scientific and research activities on the use of parametric and non-parametric quantitative methods on economic issues. He was also team member of teams dealing with experimental economics and agent-based modeling, where he participated mainly as a data miner.

Boris Kulik graduated from Leningrad Mining Institute and worked for the USSR Ministry of Geology from 1971 to 1989 in automation of drilling control. Then he took up research on logic, mathematics and artificial intelligence and got his PhD in 1996. Since 1997 Boris Kulik has worked in St.-Petersburg Institute of Problems in Machine Science of the Russian Academy of Sciences. He got his Doctor of Science (Physics and Mathematics) degree in 2008. He has published 120 scientific papers including 6 monographs.

Harendra Kumar is an Assistant Professor in the Department of Mathematics & Statistics, Gurukula Kangari Vishwavidyalaya Haridwar-Uttarakhand (India). He obtained his Master of Science in 2003 from I.I.T Roorkee and Ph.D. from Gurukula Kangari Vishwavidyalaya in 2008. He has qualified NET

(UGCJRF)- Dec.2003, NET(CSIR-JRF)-June 2005 and GATE-2004. He has published several research papers in National and International Journals in the field of Operation Research & Distributed Real Time System. He has more than 10 years teaching experience and has written three books for undergraduate & postgraduate students. He has authored several chapters for edited books. Dr. Kumar was invited to chair several international conference/workshop programs and sessions. He served in the program committees of several international conferences.

Marcus Leaning is Professor of Digital Media Education in the School of Media and Film at the University of Winchester in the UK.

Viviana Ledesma has a degree in Computer Engineer at Universidad Nacional de La Matanza in 2004. In 2019 she presented her master's thesis, which is about the definition of a mechanism for the dynamic adaptation of a requirements engineering process according to situational factors. She has participated in different Requirements Engineering research projects since 2013. She has published several papers in conference proceedings and book chapters along these years. She works as an assistant, teaching software design at Universidad Nacional de La Matanza since 2004.

Wanbil Lee is a cyberethics evangelist. He is also Principal, Wanbil & Associates; Director, Flynn Institute as well as Founder and President, The Computer Ethics Society; Co-founder & Life Fellow, The Hong Kong Computer Society; Advisory Board Member, European Centre of eCommerce & Internet Law; and Member of Nous Global International Experts Network. Dr Lee sits on several professional/learned societies and government committees. Prof Lee has held tenured and adjunct academic posts including that of Foundation Head, Department of Computing Studies, Lingnan University, Hong Kong, and Hon. Research Associate at the Computer Science Centre for Information Security & Cryptography, Hong Kong University. In addition, he held technical line and staff positions in the Australian Government and international banks. He was awarded a DBA for the dissertation on the effect of Semi-intelligent Risk-analytic Audit. His field of interest spans Computer Auditing and Ethical Computing with a current focus on techno-ethical issues: data protection & cybersecurity.

Jose Luis Lo Huang is a Computer Engineer graduated from Simón Bolívar University in Venezuela. He is specialized in Database Design and Administration, Network Services and Information Security. He has over 5 years experience in Unix / Linux Systems Administration and as instructor and developer of open source software, many of them working in the Computer Teaching Laboratory in the university. During his university period, he worked in areas like Artificial Intelligence, Robotic and Wireless Networks. Also, he was part of the university volunteer fire department, the chess group, the athletics club and co-founder of the Rubik's cube solvers club. He is member of IEEE, ACM and ISOC. Now, he is working at Oracle Corporation as Senior Consultant.

Shen Lu is a PhD candidate in Computer Science at University of South Florida. Her research interests include database systems, distributed computing, and GPU computing, especially in deep learning. Shen received BS and MS degrees in Computer Science from Tsinghua University, China. She has won Microsoft Innovation Cup Student Software Competition at the first place, and received Best Paper Awards for three times. Shen is the founder and CEO of Soft Challenge LLC which works on software

About the Contributors

development, business solutions, and web applications. At Soft Challenge, Shen has lead and conducted research for both academic and industrial customers in commercializing research products.

Sabrina Ching Yuen Luk is Assistant Professor in Public Policy and Global Affairs. She holds both Bachelor of Social Science (First Class Honours) and M.Phil from the Chinese University of Hong Kong and Ph.D. from the University of Birmingham, the United Kingdom. Dr. Luk's research interests lie broadly in aging and health care reforms, e-government and governance in Asia, digital health and cyber security, public policy analysis and China studies. She was the Highly Commended Award Winner of the 2013 Emerald/EFMD Outstanding Doctoral Research Awards in the Healthcare Management Category. She has published three books, including *Health Insurance Reforms in Asia* (New York: Routledge, 2014), *The Logic of Chinese Politics: Cores, Peripheries and Peaceful Rising* (UK: Edward Elgar Publishing Ltd, 2016) (with Peter Preston) and *Financing Health Care in China: Towards Universal Health Insurance* (New York: Routledge, 2017).

Divyaprabha M. is a student at Thiagarajar College of Engineering. Her research area include Network and Information Security.

Thangavel M. is an Assistant Professor presently working in the Department of Information Technology at Thiagarajar college of Engineering, Madurai. He presently holds 6 years of Teaching and Research experience in Thiagarajar college of Engineering, Madurai and 2 years of Teaching and Research experience in Madras Institute of Technology, Anna University - Chennai. He graduated as a B.E. Computer Science and Engineering from M.A.M College of Engineering, Trichy (Anna University - Chennai) and as an M.E. Computer Science and Engineering from J.J. College of Engineering and Technology, Trichy (Anna University - Chennai) and Pursuing his PhD from Madras Institute of Technology, Chennai under Anna University - Chennai. He is a Gold Medalist in UG and Anna University - First Rank Holder with Gold Medal in PG. His specialization is Cloud Computing, and Information Security. His Areas of Interest include DNA Cryptography, Ethical Hacking, Compiler Design, Computer Networks, Data Structures and High Performance Computing. He has published 8 articles in International Journals, 12 book chapters in International Publishers, 22 in the proceedings of International Conferences and 3 in the proceedings of national conferences /seminars. He has attended 91 Workshops / FDPs/Conferences in various Higher Learning Institutes like IIT, Anna University. He has organized 48 Workshops / FDPs /Contests/Industry based courses over the past 8 years of experience. He has been a delegate for Cyber Week 2017 organized by Tel Aviv University, Israel. He has been recognized by IIT Bombay; SAP CSR as SAP Award of Excellence with cash reward of Rs.5000/- for the best Participation in IITBombayX: FDP ICT001x Use of ICT in Education for Online and Blended Learning. He shows interest in student counseling, in motivating for better placements and in helping them design value-based life-style.

Vijayalakshmi M.N. is an Associate Professor in Department of Master of Computer Applications, RV College of Engineering®, Bengaluru, Karnataka, India. She possesses degree in M.C.A and M.Phil. She has received her Ph.D. degree in 2010 and has been honoured by ISTE, R.V.C.E chapter in 2011 for the award of PhD in Computer Science. Her area of interest is data mining and image processing. She has 22 years of teaching experience and has been honoured by Rastriya Shikshana Samiti Trust in the year 2012 for her exemplary contribution for teaching. She has been awarded as "Best Indian Citizen Gold Medal Award" by Global Economic Progress Research Association on 26th January 2016 for

Individual Achievement and National Development. She has successfully completed a research project funded by Computer Society of India as principal investigator. She has presented and published more than 55 papers at Conferences and Journals. She has received “Outstanding Scientist Award” by Venus International Foundation in 2015. She has delivered invited talks at various national and international conferences. She has guided 100 projects till now. She has 4 book chapter publications. She is member of various professional bodies.

Martin Maier received the MSc and PhD degrees (both with distinctions) from the Technical University of Berlin, Germany, in 1998 and 2003, respectively. He is a full professor with the Institut National de la Recherche Scientifique (INRS), Montreal, Canada. In the summer of 2003 he was a postdoc fellow at the Massachusetts Institute of Technology (MIT), Cambridge. He was a visiting professor at Stanford University, Stanford, from October 2006 to March 2007. Further, he was a co-recipient of the 2009 IEEE Communications Society Best Tutorial Paper Award. He was a Marie Curie IIF fellow of the European Commission from March 2014 to February 2015. In March 2017, he received the Friedrich Wilhelm Bessel Research Award from the Alexander von Humboldt (AvH) Foundation in recognition of his accomplishments in research on FiWi enhanced networks. In May 2017, he was also awarded as one of the three most promising scientists in the category “Contribution to a better society” of the Marie Skłodowska-Curie Actions (MSCA) 2017 Prize of the European Commission. He is the founder and creative director of the Optical Zeitgeist Laboratory (www.zeitgeistlab.ca). He is a senior member of the IEEE.

Abhishek Majumder received his B.E. degree in Computer Science & Engineering from National Institute of Technology, Agartala and M.Tech. degree in Information Technology from Tezpur University, Assam in 2006 and 2008, respectively. He has completed PhD from the Department of Computer Science & Engineering, Assam University, Assam in 2018. He is currently working as an Assistant Professor in the Department of Computer Science & Engineering, Tripura University (A Central University), Suryamaninagar, India. His areas of interest are Wireless Network and Cloud Computing.

Benedita Malheiro holds a PhD. and a MSc. in Electrical and Computers Engineering and a five year graduation degree in Electrical Engineering from the University of Porto (Faculty of Engineering). She is currently an Adjunct Professor at the Electrical Engineering Department of the School of Engineering - Instituto Superior de Engenharia do Porto - of the Polytechnic Institute of Porto. As a researcher, she is specialized in problems of a distributed, dynamic and decentralized nature. Her areas of interest include data science, multi-agent systems, intelligent software agents, conflict resolution, belief revision, personalisation, context-awareness, location-based services, recommendation, cloud computing and engineering education.

Theodora A. Maniou is a faculty member in the Department of Social & Political Sciences, University of Cyprus. She holds a PhD in Journalism (School of Journalism & Mass Media, Aristotle University, Greece). Her research includes a monograph regarding broadcast journalism in Cyprus and several articles in international journals regarding journalism and media studies.

Stephen W. Marshall, Ph.D., is a digital marketing and media expert. He is an award-winning professor consulting for brands such as Adobe, Coca-Cola, Texas Pete and Pal’s Sudden Service. He

About the Contributors

received his doctoral degree from the University of Florida and is Chair of the Department of Media & Communication. Dr. Marshall works for Adobe as an industry trainer/expert helping Fortune 500s around the world create amazing digital customer experiences.

Alexander Maz-Machado has a degree in Mathematics and Physics from Tolima University, a Ph.D. in Mathematical Sciences from the University of Granada. He is Associate Professor of the Department of Mathematics of the University of Córdoba, Spain.

Nigel McKelvey (Ed.D., MSc., BSc., PGCE, MICS) is a lecturer in Computing at the Letterkenny Institute of Technology and specialises in teaching secure programming techniques at both undergraduate and postgraduate level. Other areas of interest include ethical gaming, performance based programming and digital forensics. Nigel is an advocate of CoderDojos and encourages community programming initiatives.

Maria P. Michailidis holds a Bachelor degree in Business Administration-Management, two Master degrees in Community-Social Psychology and Clinical Psychology and a Doctorate degree of Education in Educational Leadership, from the University of Massachusetts, Lowell Massachusetts, USA and the University of Indianapolis, Indiana, USA. She is also a Client-Centered Psychotherapist with a Diploma in Psychotherapy of The Person-Centered Approach (PCA). Dr Michailidis is a Professor and the former Dean of the School of Business at the University of Nicosia, Cyprus and a member of the Management and MIS Department. In the past, she served as the Head of the Education Department of the University of Nicosia and Intercollege. Professor Maria Michailidis is the author of several book chapters, journal articles and editor of several books. She has a long industrial experience as she worked in the United States and Cyprus, as well as teaching experience from teaching in higher education both in the United States and Cyprus. She is a researcher and a regular invited speaker and presenter at international conferences. Most of her research work has been supported by grants; she has also been a research partner in European Union projects such as Leonardo da Vinci; Socrates/Erasmus Thematic Networks, Grundtvig Multilateral Projects and COST actions. Her research interests include occupational health, illness and wellness, gender and intercultural studies, unemployment issues, working women and stress, occupational stress, women entrepreneurs. She is a board member of the A. G. Leventis Scholars Association and also of the Elpida Foundation for Children with Cancer and Leukemia.

Branka Mikavica received her B.Sc. (2011), M.Sc. (2013) and Ph.D. (2019) degrees in telecommunication traffic engineering from the University of Belgrade – Faculty of Transport and Traffic Engineering, where she is currently assistant professor. As author or coauthor, she published more than 40 papers in national and international journals, books and conferences. Her research interests include networks integration, cloud computing and elastic optical networking.

Mohammad Izzuddin Mohammed Jamil is currently undertaking a PhD course at Universiti Brunei Darussalam. He received his Bachelor's Degree in Business Administration at Coventry University and Master's Degree in Management at Universiti Brunei Darussalam. He is an O Level and A Level Tutor, teaching Business, Accounting, Economics and Commerce. His interests include lifelong pursuit of knowledge, sharing those knowledges, involvement in the world of research and academia, and aspires to become an entrepreneur on multiple portfolios. His research interests include topics related

to Entrepreneurship, Innovation, Digital Literacy, Strategy, Organisational Theory, Investment, SMEs, Internet of Things, E-Commerce and Career Choice. He is also proficient in IT skills that is relevant for the sustainability of enterprises such as Computer-Aided Design for the designing of new products, Graphic Design to devise advertisements, Web Design for the creation and maintenance of websites, as well as Database Management.

Md. Awal Hossain Mollah is teaching at the University of Rajshahi, Bangladesh for the last 16 years in Public Administration. He is working on various issues of politics and governance in the context of Bangladesh and South Asia. He has published two chapters in “Global Encyclopedia of Public Administration, Public Policy and Governance” and one chapter in “Encyclopedia of Gerontology and Population Aging”. Besides, he has a number of peer reviewed journal articles published by Sage, Springer, Routledge, Emerald, and Palgrave including a book and some book chapters.

José Augusto Monteiro has MSc on Information Management from ‘FEUP - Faculdade de Engenharia da Universidade do Porto’ (2009). He is MCP - Microsoft Certified Professional, and has ‘CAP - Certificado de Aptidão Pedagógica’ from ‘IEFP - Instituto de Emprego e Formação Profissional’. He has started his professional career on mechatronics and industrial equipment at Cimertex SA, has served in Portuguese Air Force as an operational in Radar Maintenance area, and it has been working in IT Systems area and office automation at ‘Beltrão Coelho (Porto)’ over 10 years. Since 2004, he teaches and researches at ‘ISPGaya - Instituto Superior Politécnico Gaya’. Also, it plays the roles of the coordinator of the Course of Management of Networks and Computer Systems and it is responsible for the IT Systems that support academic activities. Now, he is a PhD candidate on Science and Web Technology at Open University in Portugal.

Ana S. Moura is a full-time researcher at the University of Porto, has authored or co-authored several books for more than a decade and takes a keen interest in such various areas as Computational Chemistry, Biomedicine, Social Sciences, and Physics.

Debapriya Mukherjee is an M.Tech student in the Dept. of Radio Physics and Electronics, University of Calcutta, India. She received her B.Tech degree in Radio Physics and Electronics in the year 2018.

Abdullah Murrar is a PhD Researcher at Indiana University of Pennsylvania in USA. Mr. Murrar has two master degrees: the first master is in Business Administration, and the second master is in Strategic Planning. He published about 10 papers in international journals. Subjects are in managing the staff who work in information technology sector, evaluating the sustainability of water sector, and developing the financial markets. Mr. Murrar has total experience about 20 years in software development management, financial management, and multidisciplinary research. He has extensive experience in management of the software development such as gathering business requirements from the end users, preparing functional design and prototyping, estimating the budgetary cost and managing the software development team. Mr. Murrar worked with international donors such as World Bank, GIZ, Cesvi, and USAID. He has experience in financial management of projects such, financial planning, financial reporting, and financial system implementation. He was speaker about financial sustainability and economic development in local radios, local and regional conferences.

About the Contributors

Ambika Nagaraj has completed her MCA in the year 2001 and M.Phil in the year 2008. She has completed her Ph.D. from Bharathiar University, Coimbatore. She has 12 years of teaching experience. She has published a significant amount of papers in chapters in books, national/ International conference papers and Journals. Presently she is working as a faculty in the computer science department at SSMRV degree college.

Rajit Nair has done a lot of research in the area of Machine Learning, IoT and Blockchain. He is working in Jagran Lakecity University as an Assistant Professor which is located at India. He has published many papers and chapters in various international journals and conferences. He is having more than 13 years of academic experience in the field of Computer Science & Engineering and also working as a reviewer of reputed International journals. He also received many awards from various organizations for contribution in the field of academics. Currently he is working on different project which are based on IoT and machine learning.

Ritu Narang is Assistant Professor Stage III at the Department of Business Administration, University of Lucknow. She has done her Ph. D. and D.Lit. in Business Administration and was first in the order of merit in Business Administration. She was awarded 'The Best Research' Award among the Teachers by the University of Lucknow in 2010. Two of her research papers have been awarded 'The Best Research Paper' in International Conferences including one organized by Oxford Business School. Twice she has been a Senior Distinguished Fellow at Hanken School of Economics, Helsinki and once at Tianjin University, China. Dr. Narang has successfully completed two Major Research Projects sponsored by University Grants Commission, New Delhi. Her papers have been presented at various national and international conferences including Academy of Marketing Science World Marketing Conference, POMS Annual Conference, Centre for Marketing in Emerging Economies (Annual Conference of the Emerging Markets Conference Board), India Academy of Management, Conference on Retailing in Asia Pacific (The Oxford Institute of Retail Management), American Marketing Association's Services Special Interest Group International Conference, etc. Dr. Narang has several publications in various international journals such as International Review of Retail, Distribution and Consumer Research, International Journal of Case Studies in Management, International Journal of Health Care Quality Assurance, Clinical Governance: An International Journal, Young Consumer, International Journal of Management in Education, Quality Assurance in Education, Vikalpa, etc. She has done Indian adaptation of a book on Organization Behaviour published by Pearson Publishers. She has published several chapters in national and international books. Dr. Narang has also been involved in delivering talks to managers of public and private sector organizations.

Sabyasachi Nayak is a Social Development Professional for the last ten years. Currently, he is working as a Counselor with the Confederation of Indian Industry. In his current position, he is extensively involved in coordinating with Government, industry and other stakeholders in addressing challenges with respect to water management.

Nicholas Nechval (Ph.D. in Automatic Control and Systems Engineering, and Dr.Habil.Sc. in Radio Engineering) is a Professor in the Department of Mathematics at the Baltic International Academy, Riga, Latvia, and Principal Investigator in the BVEF Research Institute at the University of Latvia, Riga, Latvia. Before his current position, from 1993 to 1999, Dr. Nechval was a Professor of Mathematics and

Computer Science and the Head of the Research Laboratory at the Riga Aviation University. From 1999 to 2014, Dr. Nechval was a Professor and Head (Statistics Department and EVF Research Institute) at the University of Latvia, Riga, Latvia. He was also a Principal Investigator in the Institute of Mathematics and Computer Science at the University of Latvia. Dr. Nechval, in 1991, was awarded a Silver Medal of the Exhibition Committee (Moscow, Russia) for his research on the problem of Prevention of Collisions between Aircraft and Birds. His research interests include mathematics, econometrics, stochastic processes, pattern recognition, digital radar signal processing, operations research, statistical decision theory, adaptive control, and mechanical engineering. Professor Nechval has authored and coauthored more than 455 papers and 9 books, including the book: *Aircraft Protection from Birds* (Moscow: Russian Academy of Science, 2007) coauthored with V. D. Il'yichev (Academician of the Russian Academy of Science), and the book: *Improved Decisions in Statistics* (Riga: Izglitibas Soli, 2004) coauthored with Prof. E. K. Vasermanis (University of Latvia). This book was awarded the “2004 Best Publication Award” by the Baltic Operations Research Society. Dr. Nechval is also an Editorial Board Member of several international journals.

Isaiah Oke holds a PhD degree in Civil Engineering. His main research Interest is in water Engineering. He has published a number of articles in reputable journals and Conference Proceedings. He is an academic in the Department of Civil Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.

Fehintola Olumuyiwa is a Senior Lecturer at Department at Adeyemi College of Education.

María-Dolores Olvera-Lobo is a Doctor in Documentation, Full Professor at the Department of Information and Communication at the University of Granada (Spain) and teacher at the Faculties of Communication and Documentation, and Translation and Interpreting. She has taught courses in PhD, Master's and Postgraduate programmes at various universities. She has led research groups and projects, and teaching innovation projects. She has participated in a number of international conferences. She has written and co-authored books, chapters and dozens of articles published in prestigious specialised magazines.

Aanuoluwapo Oluwaseun Omolade-Lawal is a Process Analyst at the University College Hospital, Ibadan, Nigeria. She holds a B.Sc in Computer Science from the then University of Ado-Ekiti (now Ekiti State University, Ado-Ekiti) and a Master of Information Science (M.Inf.Sc) from the Africa Regional Centre for Information Science, University of Ibadan where she is currently undertaking her doctoral research in Information Systems.

Ruslan Ostroha is a member of Processes and Equipment of Chemical and Petroleum-Refineries Department. Research areas: 1. Experimental and analytical study of hydromechanical and heat-mass transfer processes in chemical technology. 2. Studying hydrodynamics of flows with regard to their effect on thermal and physical characteristics of the suspended layer of particles. 3. Development of new types of heat-mass transfer equipment with the use of sectioning to create different conditions for the particle in the device.

Sarvesh Pandey received his Master degree from Computer Science and Engineering Department, Madan Mohan Malaviya University of Technology, Gorakhpur, India in year 2014. Presently he is

About the Contributors

working as a Senior Research Fellow (PhD Research Scholar) in Computer Science and Engineering Department, Madan Mohan Malaviya University of Technology, Gorakhpur, India. His area of interest includes Real Time Systems, Distributed Real Time Database Systems, Mobile Distributed Real Time Database Systems, Database Replication, and Cloud Service Trustworthiness. He has published more than 10 research papers in National & International conferences, Journals and Books. He is pursuing his PhD in Computer Science & Engineering. He is a reviewer of a few reputed SCI Journals.

Lambrini Papadopoulou holds a PhD in Journalism (Department of Communication, Media & Culture of Panteion University, Greece). Her thesis was concerned with new business models for journalism. She currently teaches Web journalism in Panteion University while practicing journalism in Greece. Her research has been published, among others, in Journalism Studies and Journalism Practice.

Angeliki Papaioannou studied Physics and Computer Science at the National and Kapodistrian University of Athens and received her B.Sc. in Physics in 2003 and her M.Sc. in Electronics and Automation with specialization in Informatics and Information Systems in 2010. Angeliki has worked for LogicDIS as Software Engineer and later as Product Manager of Securities ERP. In 2017, she joined as an associate software engineer INFOLYSiS, developing various IoT applications of INFOLYSiS portfolio, using Arduino nodes, Zigbee, web, mobile and Java tools. Her current research interests include interoperability IoT mapping functions for heterogeneous IoT platforms, SDN, NFV, Cloud Computing, IoT infrastructures, Monitoring systems and web/mobile applications.

Satyam Paul is a researcher in the Department of Mechanical Engineering, Örebro University. He earned his PhD from the Department of Automatic Control, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico. He has a Master degree in the field of Mechatronics from VIT University, India and a Bachelor degree in Mechanical Engineering from National Institute of Technology, India. His research area includes control systems, intelligent techniques, vibration control and predictive maintenance.

Yuri Pavlov is associate professor in the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. He received DUES from Paris VI - France, MSc degree in Automation from Technical University of Sofia and holds PhDs from the BAS. He has publications in: Proceedings of Bulgarian Academy of Sciences, Bioautomation, EJOR, E-learning III and the knowledge society-Belgium, Proceedings in Manufacturing Systems – Romania, Global Journals Inc. (USA), IOSR JCE. He is author of the monograph “Decision Control, Management, and Support in Adaptive and Complex Systems: Quantitative Models. IGI Global and several chapters in “Encyclopedia of Business Analytics and Optimization”, Encyclopedia of Information Science and Technology, Fourth Edition (IGI Global) “Artificial Intelligence: Concepts, Methodologies, Tools and Applications” - IRMA USA. He has interests in the fields of decision making, utility theory, theory of measurement, stochastic approximation, machine learning, optimization and control, approaches for control, reduction and equivalent transformation of nonlinear models. He is member of International Institute of Informatics and Systemics (IIS).

Carla Pires is Professor at Universidade Lusófona's Research Center for Biosciences and Health Technologies, Portugal, and invited collaborator Faculty of Arts and Humanities of the University of

Lisbon, and Laboratory of Phonetics and Lisbon Baby Lab. Her research interests include health communication, social pharmacy, pharmacy practice, and other. <http://orcid.org/0000-0003-4526-0271>.

Julia Margarete Puaschunder studied Philosophy/Psychology (MPhil, University of Vienna, 2003), Business (MBA, Vienna University of Economics and Business, 2007), Public Administration (MPA, Maxwell School, 2008), Social and Economic Sciences (Doctor, Vienna University of Economics and Business, 2006), Natural Sciences (Doctor, University of Vienna, 2010), Law and Economics (pending). Julia M. Puaschunder has launched and administered research projects in Australia, Austria, Canada, China, Germany, Indonesia, Switzerland, and the United States. She conducted interdisciplinary science and governance projects as well as held or holds advisory, board member, evaluator, committee, contributor, observer, referee and representative privileges for the Academic Council on the United Nations System, Academy of Behavioral Finance and Economics, Alpbach Laxenburg Group, Austrian Academy of Sciences, Austrian Chamber of Commerce New York, Austrian Economic Association, Austrian Federal Ministry of Science and Research, Austrian Federal Ministry of Science, Research and Economy, Austrian General Consulate New York, Austrian Office of Science and Technology Washington D.C., Austrian Scientists in North America Association, Austrian Society for New Institutional Economics, Cambridge Scholars Publishing, Discovery Innovation Growth (DiG) Talent Development Network, EURO working group on Behavioural Operational Research, European Horizons Conference, European Union Commission, Finance, Risk and Accounting Perspectives Conference at University of Cambridge, Fritz Thyssen Foundation, Fulbright Commission, Harvard Association for Law and Mind Sciences, Harvard Club of Austria, Harvard University Center for the Environment, Horizon 2020, International Institute for Applied Systems Analysis, Janeway Center, Kollegium Kalksburg Altkalksburger Vereinigung, Netherlands Organisation for Scientific Research, Oikos International New York City Chapter, Oxford Academic Research Network, Research in Management and Learning Education Unconference, Society for International Development Vienna Chapter, Society of Authors, The Max Kade Foundation New York, The Neue Galerie New York, Tishman Environment and Design Center, the U.S. Department of Education, Vienna Foundation, Vienna International Conference on Mathematical Modelling and Viennese Opera Ball New York International Business Committee. In 2011, she served as a participant in a U.S. White House conference call on environmental justice. Before starting a Prize Fellowship in the Inter-University Consortium of New York working for the New School (Deans Office, Department of Economics, Department of Politics, Economics of Climate Change Project, Eugene Lang College, Parsons School of Design, School of Public Engagement and the Schwartz Center for Economic Policy Analysis) as well as Columbia University (Department of Economics, Graduate School of Arts and Sciences), Julia M. Puaschunder held positions at the University of Vienna and the Vienna University of Economics and Business. For several years she was an Associate of the Harvard University Faculty of Arts and Sciences and serves as Contributor to the Harvard Law School Law and Mind Sciences Initiative Situationist. At the Harvard University Center for the Environment, she conducted research on intergenerational equity in cooperation with Harvard Business School and Harvard Kennedy School. Trained as a behavioral economist with Doctorates in Social and Economic Sciences as well as Natural Sciences and Masters in Business, Public Administration and Philosophy/Psychology, she has over 15 years of experience in applied social sciences empirical research in the international arena. Julia M. Puaschunder serves as expert member of the Academic Council on the United Nations System, the European Corporate Governance Institute, and the Virtus Corporate Governance Experts Global Repository. Apart from academic appointments in Austria, Julia was a scholar at The Australian National University, Columbia University in the

About the Contributors

City of New York, Harvard University, Haskayne School of Business, Ludwig-Maximilians-University of Munich, and The Open Society Institute & Soros Foundation New York. Currently, she conducts research as a Prize Fellow in the Inter-University Consortium of New York. Throughout her academic career, Julia was invited to present her research at Harvard University, Princeton University, Columbia University, Brown University, Oxford University and Cambridge University as well as The Academic Council on the United Nations System. She published with Harvard University, Columbia University and Oxford University outlets among other distinct journals and international publishing houses (e.g., Emerald, Palgrave Macmillan, Routledge, Springer International & Nature, Vernon, Wiley). Julia organized conferences and public speakers from the Harvard Business School and Princeton University in North America and Europe. Since 2015 Julia Puaschunder supports the Economics of Climate Change Project Speaker Series hosted at The Schwartz Center for Economic Policy Analysis of The New School for Social Research in New York City. Julia helped initiate the 2017 Unconference at The New School New York in cooperation with the New School for Public Engagement and served on the Scientific Committee of the 2017 Finance, Risk and Accounting Perspectives Conference as well as the Social and Sustainable Finance and Impact Investing Conference 2017 at University of Cambridge. Julia is scheduled as scientific board member of the 2018 Vienna International Conference on Mathematical Modelling at the Technical University Vienna as well as the 2019 Annual Conference of the Academy of Behavioral Finance and Economics. Julia is included in the '2018 Marquis Who's Who in America and in the World' among the top 3% professionals around the globe. She was awarded the 2018 Albert Nelson Marquis Lifetime Achievement Award. After having captured social responsibility in corporate and financial markets in Europe and North America with attention to Financial Social Responsibility and Socially Responsible Investment; Julia Puaschunder currently pursues the idea of Eternal Equity — responding to Western world intergenerational equity constraints in the domains of environmental sustainability, overindebtedness and demographic aging with focus on law and mind sciences. Currently, Julia dedicates interest to Humanness, Automated Control and Artificial Intelligence Ethics.

Kornelije Rabuzin is a full professor of databases at the Faculty of organization and informatics, University of Zagreb, Croatia. He holds Bachelor, Master, and PhD degrees - all in Information Science. He performs research in the area of databases, data warehouses, business intelligence, and data science. He is the head of the department for theoretical and applied foundations of information science.

Monica Bixby Radu is an Assistant Professor of Sociology at Southeast Missouri State University in the Department of Criminal Justice, Social Work, and Sociology. Some of her current research interests include intimate partner violence, school safety, bullying, family and school social capital, and juvenile delinquency. Dr. Radu's work has been published in *American Behavioral Scientist*, *Child Development Perspectives*, *Sociology Compass*, and *The Journal of Public and Professional Sociology* among other scholarly journals and edited books.

Bapuji Rao is currently working as Asst. Professor of Computer Science, Engg. & Applications at Indira Gandhi Institute of Technology (IGIT), Sarang, Dhenkanal, Odisha, India. He received M.Tech (Computer Science) from Berhampur University, Berhampur, Odisha, India. He is pursuing Ph.D. (CSE) from Biju Patnaik University of Technology (BPUT), Rourkela, Odisha, India. He has published 27 papers in International Conferences which includes IEEE-6, Springer-4, Elsevier-1, and McGraw-Hill-1. He has published 8 papers in International Journals of repute. He has published four book chapters in

IGI Global, USA. His Current Research area focuses on Graph Mining, Social Network, Data Mining, Multi-Layer Graph, and attributed graph.

Jose Poças Rascão (PhD, ISCTE/ULI) is a Professor of Management Information Systems at the Polytechnic Institute of Setúbal, Portugal. It is research in the Research Unit in the development of enterprise (UNIDE) at the Lisbon University Institute (ISCTE/ULI). Since 2000 has published 16 books and chapters of books on strategy management and management of information and more than 43 publications on strategy and information management. Under his supervision, 17 Masters Theses have been successfully from 2007 to 2018, with 3 more coming up for defense in next 2 years. From 2010 to 2015 was Director of course management of distribution and logistics. He also works the private management consultant for large and medium-sized enterprises. He worked the management consult at Price Waterhouse and Coopers & Lybrand between 1989 and 1995.

Adnan Rashid received his PhD degree in Information Technology from School of Electrical Engineering and Computer Science (SEECS), National University of Science and Technology (NUST), Islamabad, Pakistan, 2019. Prior to this, he received the M.Sc. and M.Phil. degree in Electronics from the Department of Electronics, Quaid-i-Azam University (QAU), Islamabad, Pakistan, in 2008 and 2012, respectively. Currently, he is working as a Research Associate with the System Analysis and Verification (SAVE) laboratory in SEECS, NUST. He also worked as a Visiting Researcher at Hardware Verification Group (HVG), Concordia University, Canada, for 6 months in 2017. He has a strong interest in Formal Methods, with their applications in Control Systems, Analog Circuits, Biological Systems, Communication Systems and Transportation Systems. He has served as a chair of the Doctoral program at Conference on Intelligent Computer Mathematics (CICM-2017), Edinburgh, UK.

Mamata Rath, M.Tech, Ph.D. (Comp.Sc), has twelve years of experience in teaching as well as in research and her research interests include Mobile Adhoc Networks, Internet of Things, Ubiquitous Computing, VANET and Computer Security.

Maya Rathore has received her Master Degree in Computer Applications from Rajiv Gandhi Technical University, Bhopal, India and PhD in Computer Science from Devi Ahilya University Indore, India. Presently she is working as Associate Professor at Christian Eminent College, Indore (MP), India. She is having 12.4 years of teaching and 8 years of research experience. Her areas of research are Software Engineering, Service- Oriented Computing, Distributed computing and Cloud computing. She has contributed 18 research publications in various national and international conferences and journals. She has co-authored a book on OOAD and published 1 book chapters in Springer.

Sina Razvarz received the B.S. degree in Mechanical Engineering, Heat and Fluids from Islamic Azad in Bonab, Iran in 2009 and the M.S. degree in Mechanical Engineering from Islamic Azad University, Science and Research Branch, Tehran, in 2010. Now, He is studying at the Departamento de Control Automático, CINVESTAV-IPN (National Polytechnic Institute), Mexico City, Mexico, for his Ph.D. degree. His research interests include heat transfer, fluid dynamic, fuzzy equation, neural networks, and nonlinear system modeling.

Irene Rivera-Trigueros is a PhD student and research staff in training at the University of Granada.

About the Contributors

James A. Rodger is a Professor of Management Information Systems and Decision Sciences at Indiana University of Pennsylvania (IUP). He received his Doctorate in MIS from Southern Illinois University at Carbondale in 1997. Dr. Rodger has published several journal articles related to these subjects. His work has appeared in the following journals: Computers in Human Behavior, Journal of Computer Information Systems, Issues in Information Systems, IEEE Transactions on Software Engineering, International Journal of Hybrid Intelligent Systems, Information and Software Technology, Information Technology and Management, Annals of Operations Research, Communications of ACM, Computers & Operations Research, Decision Support Systems, Expert Systems with Applications, Lecture Notes in Computer Science, International Journal of Human-Computer Studies as well as several other journals.

Alexandria L. Rook is an undergraduate student at Southeast Missouri State University majoring in Criminal Justice and minoring in Spanish, with special interests in corrections and gender studies. She is interested in pursuing a career in law enforcement and has research interests in cyberbullying, racial disparities within the prison system, and policy development.

Sudhir K. Routray received his B.E. degree in electrical engineering from Utkal University, India; M.Sc. (Engineering) degree in data communications from Sheffield University, UK; and PhD in telecommunication engineering from University of Aveiro, Portugal. He has worked as a lecturer in Biju Pattnaik Technical University, Rourkela (India); Swami Vivekananda University of Technology, Bhilai (India); and Eritrea Institute of Technology, Mai Nefhi (Eritrea). He joined the Institute of Telecommunications, Aveiro in 2011 and worked there till 2015. He joined CMR Institute of Technology, Bangalore, India in Aug 2015 as a member of the faculty of Electronics and Communication Engineering. His research interests are optical networks, network science and statistical analysis of communication networks. He is a senior member of IEEE. He serves as a volunteer in IEEE Bangalore Section in IEEE Region 10. Now he works as an Associate Professor in the department of ECE in AASTU.

Khadijeh (Roya) Rouzbehani, PhD, is a Postdoctoral Fellow with teaching and research interests in policy, Collaboration, strategic communication, Public Health, Knowledge Management, and corporate social responsibility (which includes ethics and sustainability). She started teaching at the School of Management of UT in 2012, while running her own consulting center, as well. She has developed and taught courses for the MPA, and Business Administration programs. Her first book was on Social and Economical aspects of Public Health Policy and is getting the second one published which is on political factors influencing Public Health. She joined University of Victoria as a Postdoctoral fellow in 2017 to further her research and teaching in this top comprehensive university.

Nagayuki Saito is a former Policy Analyst of OECD, Directorate for Science, Technology and Industry, Special Appointment Senior Research Fellow of Ministry of Internal Affairs and Communications, Institution for Information and Communication Policy and visiting researcher of Aoyama-Gakuin University Human Innovation Research Center. He is instructor of Ochanomizu University Interdisciplinary liberal arts. He graduated from a Doctorate Program of Keio University Graduate School of Media Design and had a Master's degree of Economics at Chuo University Graduate School. He is a committee Member of the Committee for Information, Communications and Consumer Policy Division, Working Party on Information Security and Privacy in OECD. His research theme is making a safer Internet environment

for the youth. It focuses on awareness of education, making collaborative e-learning system based on theory of social constructivism; information and communication policy.

Christos Sakkas received his BSc degree in Informatics and his MSc degree in Computer Science from Department of Informatics of Athens University of Economics and Business (AUEB). Mr. Sakkas has been an active research associate of Media Net Lab at the NCSR “Demokritos” from 2013 until 2018. Since 2018 he has joined the R&D team of INFOLYSiS P.C. He has participated in many EU-funded projects, collaborating with various universities and enterprises with presentations and publications at international conferences. His current research interests include IoT, SDN, NFV, Cloud Computing, Monitoring systems and web/mobile applications and web/mobile applications including chatbots.

Wilson Sakpere received his bachelor’s degree in Electrical Engineering in 2007 from the University of Ilorin, Nigeria and master’s degree in Information Technology in 2015 from the Cape Peninsula University of Technology (CPUT), Cape Town, South Africa. He is a quality-focused researcher with a balanced background between academic and practical expertise. He is presently a doctoral researcher at the Computer Science and Engineering department, Alma Mater Studiorum, University of Bologna, Italy. His research interests include wireless networks, digital technologies, computer vision, localisation, navigation and wayfinding. The interests are directed towards applying innovations in a social context and application. As a student, engineer and lecturer, he has served in various leadership roles and engaged in community service. Hence, he has been awarded a couple of awards and commendations.

Antonio Sarasa Cabezuelo has a MS. degree on Mathematics of Computer Sciences from the Complutense University of Madrid, BS. degree on Computer Sciences from the National University of Distance Education, MS. degree on Computer Sciences from the Open University of Catalonia and CS PhD from the Complutense University of Madrid. Currently, he is an associate professor in the Computer Science School at Complutense University of Madrid, and a member of the research group ILSA (Implementation of Language-Driven Software and Applications,, <http://ilsa.fdi.ucm.es>). His research is focused on e-Learning, markup languages and Domain-specific languages. He was one of the developers of the Agrega project on digital repositories. He has published over 50 research papers in national and international conferences. Likewise, he is a member of the 36 Subcommittee of AENOR.

Theodora Saridou received her BSc in Law and her MSc in Journalism and New Media, both from Aristotle University of Thessaloniki. Her master thesis dealt with tools and applications of participatory journalism, studying, among others, the legal and ethical issues stemming from it. She is currently a PhD candidate in the School of Journalism and Mass Media Communication, researching the changing conditions in online media due to citizens’ participation in the news production process. Her research interests include participatory journalism, user-generated content, new media and online news production.

Rui Sarmento has a degree in Electrical Engineering in the Faculty of Engineering, University of Porto, and an MSc in Data Analysis and Decision Support Systems in the Faculty of Economics of the University of Porto. He has worked in several areas from an international technical support center to software development companies focusing on Communications and Intranet solutions with Linux-based Enterprise Operating Systems. He has also worked for the leading public transportation company in his hometown, Porto, as a Project Management engineer in the IT area. Rui Sarmento is also involved

About the Contributors

with the Laboratory of Artificial Intelligence and Decision Support at INESC TEC, researching Large Social Networks Analysis and Visualization for his Ph. D. thesis. He is currently also collaborating with the Statsframe team (Statistical Services and Software). Rui was previously published as a Contributing Author in the book “Integration of Data Mining in Business Intelligence Systems” (IGI Global, 2014), and he also contributed with the book “Comparative Approaches to Using R and Python for Statistical Data Analysis” (IGI Global, 2017).

Gülcan Sarp is a Geographic Information System and Remote Sensing specialist with the research interests in active tectonics and geomorphology. She obtained a B.S. in geology from Ankara University in 1998, M.S. and Ph.D. in Geodesy and Geographical Information Technologies from Middle East Technical University in 2005 and 2012 respectively. She is currently working at Süleyman Demirel University as a Associated Prof. Dr.

Richard S. Segall is Professor of Computer & Information Technology in Neil Griffin College of Business at Arkansas State University. He holds BS/MS in mathematics, MS in operations research and statistics from Rensselaer Polytechnic Institute in Troy, New York, and PhD in operations research from University of Massachusetts at Amherst. He has served on faculty of Texas Tech University, University of Louisville, University of New Hampshire, University of Massachusetts-Lowell, and West Virginia University. His research interests include data mining, Big Data, text mining, web mining, database management, and mathematical modeling. His funded research includes that by U.S. Air Force, NASA, Arkansas Biosciences Institute (ABI), and Arkansas Science & Technology Authority (ASTA), and is member of the Arkansas Center for Plant-Powered-Production (P3), Editorial Board of International Journal of Data Mining, Modelling and Management (IJDM), International Journal of Data Science (IJDS), and co-editor of Handbook of Big Data Storage and Visualization Techniques; Research and Applications in Global Supercomputing; Visual Analytics of Interactive Technologies: Applications to Data, Text & Web Mining and Open Source Software for Statistical Analysis of Big Data; all published by IGI Global.

Devjani Sen received her Doctorate, Master’s and Bachelor’s (class of Highest Honors) degrees in Psychology from Carleton University. She has nearly two decades of experience working as a social psychologist and approximately ten years of experience in human factors design and research. She has published in the areas of media, healthcare, information and communication technologies, public health crisis management, privacy, and information credibility and presented the results of her research at conferences in North America, Europe and Asia. Her work in human information technologies began in the early 2000s. Since then, she has been involved in a number of industry projects including current research trends using digital platforms. These projects comprised field research and testing of various decision support tools in the areas of healthcare, learning, human resources, and criminal analysis that best meet the needs of end-users. Numerous tools she has helped to develop have been used Canada-wide by the Department of Defence and First Responders, providing access to databases, interactive maps, checklists, incident reports, emergency response protocols. These tools have also been used by fire, emergency response, defence, and law enforcement agencies.

Mirjam Sepesy Maučec received the B.Sc. and Ph.D. degrees in computer science from the Faculty of Electrical Engineering and Computer Science, University of Maribor, in 1996 and 2001, respectively.

She was a Researcher with the Center for Interdisciplinary and Multidisciplinary Research and Studies, University of Maribor, from 1996 to 1997, and a young Researcher with the Faculty of Electrical Engineering and Computer Science, University of Maribor, under the guidance of mentor Prof. Dr. Z. Kačič, from 1997 to 2001. In 2010, she became an Associate Professor in telecommunications. She has been involved in several projects and in national research program Advanced Methods of Interaction in Telecommunication. She participated in several international and national research projects. Her research interests include machine translation, computational linguistics, and evolutionary computing.

Udai Shanker is presently Professor in the Department of Computer Sc. & Engineering of M. M. M. University of Technology, Gorakhpur-273010. For his imitation of the most modern of approaches and also for his exemplary devotion to the field of teaching, and sharing his profound knowledge with students to make better future citizen of India, he has been a role model for the new generation of academicians. Besides introduced radical and revolutionary changes that have positively impacted the database world and student community, he is a man well versed with all the intricacies of academics.

Paawan Sharma is an Asst. Professor in Dept. of ICT at PDPU. He completed Ph.D. in Engineering from Homi Bhabha National Institute, Mumbai. He is Author/co-Author of several publications in reputed International journals. He has also guided/co-guided 2 Ph.D. students in the area of Smart grid technology and Embedded Systems for Disaster management. He also has 3 Indian patents published.

Tetiana Shmelova is Doctor of Science, Professor of Department of Air Navigation Systems in National Aviation University (Ukraine). Areas of Scientific Interests: Artificial Intelligence; Mathematical Models of Decision-Making by operators in the Air Navigation System, especially in Emergency; Decision Support Systems; Research of Air Navigation system as Socio-technical system; Professional Effectiveness using Aviation Sociometry and Socionics; Management and Marketing in aviation; Information technology and Informatics of Decision-Making in the Air Navigation System; Problems of Human Factors in aviation. Author of more than 300 scientific articles, books, methodical manuals, copyright certificates, guides, monographs in fields of aviation, economics, mathematics, the theory of system. Teaching courses: Decision Making; Mathematical Programming; Effectiveness of Air Traffic Management; Effectiveness of Unmanned Systems; Artificial Intelligence.

Vaibhav Shukla has received M.Sc. in Mathematics from CSJM University Kanpur (UP), India in 2005, and the MCA degree from Kamla Nehru Institute of Technology Sultanpur (UP), India in 2009. He is currently pursuing Ph.D. in Computer Science and Engineering from Birla Institute of Technology Mesra Ranchi (Jharkhand), India. Currently, he is working as a Assistant Professor in the computer science department of Allenhouse Institute of Technology Kanpur, India. His research area includes optical packet switching design and analysis and computer networks.

Yuliya Sikirda is Associate Professor of Management, Economy, Law and Tourism Department at the Flight Academy of National Aviation University. Has graduated from the Flight Academy of National Aviation University (former State Flight Academy of Ukraine) in 2001, received a Master degree on speciality “Air Traffic Services” specialization “Air Traffic Control” (2001). Has got a diploma of PhD in Technical Sciences on speciality “Automated Control Systems and Advanced Information Technologies” (2004). PhD thesis “Modelling of Decision Support System for an Air Traffic Controller in Flight

About the Contributors

Emergency Situations”, thesis has defended in Kyiv Automation Institute. In 2006 has received an attestat of Associate Professor of Management and Economy Department. Since 2014 is the member of the specialized scientific council for the defence of dissertations for PhD in Technical Sciences on the speciality “Navigation and Traffic Control”. The author and co-author of about 180 scientific works, including three manuscripts and more than 40 articles in specialized scientific publications, three copy-right certificates on computer programs. Together with the scientific work has actively engaged in the methodological provision of the educational process, during the years of labour activity above 70 teaching and methodological works have been issued. Areas of scientific interests: increasing of the decision making efficiency by the human-operator of the Air Navigation System in flight emergencies; assessment of the influence of aviation enterprises’ internal and external management environment factors on the activity of the Socio-technical System’s operator. Teaching courses: ”Introduction to the Speciality”, “Transport and Transport Infrastructure”, “Information Systems in Management”, “Methodology of Scientific Researches”, ”Mathematical Modelling of Professional Tasks”.

Udit Singhania is right now pursuing BTech in Computer Science at VIT University, Vellore. He is involved in the research of the fields like Artificial intelligence, machine learning, and data science.

Roberta Spallone is Full Professor of Drawing at the Politecnico di Torino, Department of Architecture and Design (DAD). Graduated in Architecture at the Politecnico di Torino and obtained a Ph.D. in “Drawing and Survey of Architectural Heritage” at the Università “La Sapienza” di Roma. Currently she teaches in the “Laboratory of Drawing and Survey” and “Digital techniques of representation” in the Architecture Laurea. She also teaches a class of “Drawing and modeling aimed to analysis, interpretation and virtual reconstruction of architectural and urban heritage” in the Ph.D. program of Architectural and Landscape Heritage. Member of the Board of the Doctorate of Architectural and Landscape Heritage. Director with Marco Vitali of the book series “Drawing and Vaulted System”, Aracne Editrice. Member of the Scientific Committee of International Journal DisegnareCON and Revista EGE, and the Editorial Board of the journal Archaeology & Anthropology, Crimson Publisher and International Journal of Environmental Research and Public Health. Reviewer for many international magazines, peer-reviewed books and international congresses. Her research focuses on the fields of: history and criticism of architectural drawing, with particular attention to architectural treatises, urban survey, and digital techniques of representation aimed to graphical analyses and reconstructive modeling. Her Ph.D. thesis was awarded with the Silver Plaque, 1° Prize by Unione Italiana Disegno. Best paper FORT-MED2015_Valencia, Topic: Historical Research. Best paper UID2018_Milan, Topic: Techniques for Analysis and (Re)Building of Heritage.

G.R Sridhar MD DM FACE FRCP is the Director of Endocrine and Diabetes Centre Visakhapatnam and Adjunct Professor, Bioinformatics at the AU College of Engineering. The Editor in Chief, International Journal of Diabetes in Developing Countries, he is the Executive Editor of RSSDI Textbook of Diabetes, and is an editor of the ESI Handbook of Endocrinology. The founding editor of Indian Journal of Endocrinology and Metabolism he was the President of Research Society for the Study of Diabetes in India (2010). Dr Sridhar’s major areas of research interest are: Clinical informatics, computational biology and bioinformatics, psychosocial aspects of diabetes.

Ioannis Stergiou holds a BSc degree in Agriculture Engineering from the Aristotle University of Thessaloniki, Greece and a MSc degree in Agriculture and Environmental Sciences from the University of the Aegean, Greece. Since 2010, Ioannis has worked as agronomist and technical support IT officer at various agricultural commerce trade and IT related companies with focus on GIS and IoT smart farming through the use of IT systems for monitoring crop fields with the help of IoT sensors (light, humidity, temperature, soil moisture, etc.) and automating the irrigation systems and methodologies.

Ugrasen Suman has received PhD Degree in Computer Science from Devi Ahilya University Indore, and his Master Degree in Computer Applications from Rani Durgawati University Jabalpur India. Presently he is working as a Professor at School of Computer Science & Information. He has authored three books Software Engineering: Concepts & Practices (Cengage Learning, 2013), Software Engineering (Cengage Learning, 2018), Object Oriented Analysis & Design using UML (Cengage Learning, 2018). He has published three book chapters and 90 research papers in National & International Journals/ Conferences. He has more than 19 years of teaching and research experience.

Apostolos Syropoulos holds a B.Sc. in Physics, an M.Sc. in Computer Science, and a Ph.D. in Theoretical Computer Science. His research interests focus on computability theory, category theory, fuzzy set theory, and digital typography. He has authored or co-authored nine books and over 60 papers and articles.

Khaled Tamzini received his Ph.D. degree with honors from the Faculty of Economic Sciences and Management Sciences of Tunis in 2014. He justifies an experience of over 10 years (since 2001) as a senior trainer for the account of several national companies in the field of Human Resources Management. Dr. Tamzini is actually an Assistant professor of Strategic Human Resource Management at the IHEC of Sousse (Tunisia) where he teaches Human Resource Management and Strategic management. He is a panel member of the International Association of Strategic Management (France). He is a reviewer for several international scientific meeting as the Academy of Management Meeting, the Academy of International Business, the European Academy of Management Conference... His research interests include the Resource-Based View, Knowledge Management, Talent management, Employer branding, Human resource marketing, Human Capital, Entrepreneurship, Sustained Competitive Advantage and Strategic Human Resource Management.

George Theodoropoulos holds a BSc degree in Mathematics from the University of Athens, Greece. George has more than thirty years professional experience as computer analyst and programmer in the shipping sector participating in various research and development IT projects related to maritime software, vessel communications as well as vessel data transfer and data synchronization with offshore offices. Currently, he is a senior research and development engineer dealing with the applicability of IoT in the shipping sector.

Sonal Tiwari is a PhD candidate in the Department of Business Administration, University of Lucknow, India. She is UGC-NET/JRF (Junior Research Fellowship) qualified. Her research focuses on consumer behaviour, digital media consumption and marketing. She holds an MBA degree in Human Resources and Industrial Relations from the University of Lucknow.

About the Contributors

B. K. Tripathy is now working as a Dean of SITE Department VIT University, Vellore, India. He has received research/academic fellowships from UGC, DST, SERC and DOE of Govt. of India. Dr. Tripathy has published more than 430 technical papers in international journals, proceedings of international conferences and edited research volumes. He has produced 26 PhDs, 13 MPhils and 4 M.S (By research) under his supervision. Dr. Tripathy has published two text books on Soft Computing and Computer Graphics. Dr. Tripathy has served as the member of Advisory board or Technical Programme Committee member of several International conferences inside India and abroad. Also, he has edited two research volumes for IGI publications. He is a life/senior member of IEEE, ACM, IRSS, CSI, ACEEE, OMS and IMS. Dr. Tripathy is an editorial board member/reviewer of more than 60 journals.

George Tzanis holds a BSc, a MSc, and a PhD degree in Computer Science from the Aristotle University of Thessaloniki, Greece. He is also a family physician/general practitioner, holding a MD and a MSc in Primary Health Care from the Aristotle University of Thessaloniki.

Cyprian Ugwu is currently with the University of Nigeria, Nsukka and a Research Fellow with the University of South Africa. He has over 50 research publications in high impact international and local journals in Library and Information Science. Dr. Ugwu has made presentations at professional conferences locally and internationally. His research interests include knowledge management, ICT applications in libraries, digital communication and scholarship, research methodology and knowledge organization. He is happily married with children and his hobbies include reading, solving calculus and playing scrabble.

Immanuel Ovemeso Umukoro is a Research Fellow at the Sustainable and Inclusive Digital Financial Services initiative of the Lagos Business School, Pan-Atlantic University, Nigeria. He is also a doctoral student in information systems at the Africa Regional Centre for Information Science (ARCIS), University of Ibadan, Nigeria where he obtained a Master of Information Science degree. His Bachelor of Science degree is in Library and information science from the Delta State University, Abraka, Nigeria. He is a chartered member of the Nigerian Institute of Management and an Associate of the Centre for Global Enterprise (CGE) with a distinction in Platform Management, Strategy, and Innovation. Immanuel is a certified Digital Money Expert by the Fletcher School of Law and Diplomacy, Tufts University, USA. He also holds a certificate in interoperable and inclusive payment systems, digital financial services in emerging markets from the Digital Frontiers Institute as well as a certificate in research and grant proposal writing from the INASPIAuthorAID Foundation, USA.

Kiminori Usuki is a researcher at Center for Research on Educational Testing.

Cristóbal Vargas Jarillo received the B.S. degree from the Instituto Politécnico Nacional, Mexico, and the M.S. degree in Computer Science from University of Wisconsin, USA, and the Ph.D. degree in Mathematical Sciences from the University of Texas-Arlington. He works at Departamento de Control Automatico -CINVESTAV-IPN, México.

Rafael Vargas-Bernal received a Bachelor degree in Communications and Electronics Engineering from the University of Guanajuato in 1995, and the degrees of Master of Science and Doctorate in Science with Specialty in Electronics from the National Institute of Astrophysics, Optics and Electronics (INAOE) in 1997 and 2000, respectively. Since January 2002, he has been a professor-researcher

at the Higher Technological Institute of Irapuato (ITESI), Mexico and, particularly since 2006, he has worked in the Department of Materials Engineering where he has established himself as a researcher. He has authored 15 articles in journals, 26 chapters in books and about 100 conference articles. He is a member of the National System of Researchers (SNI-Mexico). He regularly serves as a reviewer of scientific articles in RSC Advances and Royal Society Open Science as well as Standards in Semiconductor Equipment and Materials International (SEMI). His research interests include two-dimensional materials, nanomaterials, space materials, composite materials, gas sensors and biosensors.

Andreas Veglis is Professor, head of the Media Informatics Lab at the School of Journalism & Mass Communication at the Aristotle University of Thessaloniki. He received his BSc in Physics, MSc in Electronics and Communications, and PhD in Computer Science, all from Aristotle University. His research interests include information technology in journalism, new media, course support environments, data journalism, open data and distance learning. He has published in books and international journals. He is the author or co-author of 12 books, he has published 74 papers on scientific journals and he has presented 125 papers in international and national Conferences. Professor Veglis has been involved in 30 national and international research projects.

Bruno Veloso holds a M.Sc. in Electrical and Computers Engineering, Major in Telecommunications from the Polytechnic Institute of Porto (School of Engineering) and a Ph.D. degree in Telematics from the University of Vigo, Spain. He is a researcher at the INESC TEC (Laboratory of Artificial Intelligence and Decision Support (LIAAD)). He is currently working at University of Portucalense as an Assistant Professor and he is also a collaborator at REMIT. His interests include distributed artificial intelligence, multi-agent systems, personalisation, recommendation systems and data streams.

Mário Véstias received the electrical engineering degree in 1993, and the PhD in electrical and computer engineering in 2002, both from the Technical University of Lisbon, Portugal. He is a Coordinate Professor at the Polytechnic Institute of Lisbon, School of Engineering (ISEL), Department of Electronic, Telecommunications and Computer Engineering (DEETC), where he is responsible for undergraduate and graduate courses on computer architecture and digital systems design. He is also a senior researcher at the ESDA (Electronic Systems Design and Automation) group at the research institute INESC-ID in Lisbon. His current research interests are about architectures for on-chip many-core systems for high-performance reconfigurable computing and for high-performance embedded computing and computer arithmetic.

Damjan Vljaj graduated from the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia, in 1998. He was awarded a Ph.D. in 2007 by the same Faculty, where he is currently employed as an Assistant Professor in the field of Telecommunications. His research interests are in the areas of speech and language technologies. He focused specifically on the research of digital signal processing in robust speech recognition in the intelligent environments and environments of fixed and mobile telephony.

Michael Gr. Voskoglou is an Emeritus Professor of Mathematical Sciences at the School of Technological Applications of the Graduate T.E.I. of Western Greece. He is the author of 17 books and of more than 500 papers published in international mathematical journals and conference proceedings.

About the Contributors

Steven Walczak, Ph.D., is an Associate Professor of Health Information Science in the School of Information and the Florida Center for Cybersecurity at the University of South Florida. His primary research interest is in applied artificial intelligence, including artificial neural networks, other machine learning methods, expert systems, heuristics, and cognitive modeling and tends to focus on healthcare related real-world problems.

Andrew Weiss is a Digital Services Librarian at California State University, Northridge.

Immo H. Wernicke served to the German Government in different positions as economic expert, political consultant, statistician and telecommunication officer. Degrees from the Faculty of Social and Economic Sciences of the University of Cologne. IT supported research and thesis on comparison of National elections in UK, France, Germany and Norway. Sales manager in a German Dax-Enterprise. Charged by the German Federal Statistical Office to implement EU- and World Bank Projects of the Acquis Communautaire on Agriculture for Turkey, Tajikistan, Zambia, South Africa, Poland and Lithuania. Charged by the German Privatization Agency (Treuhandanstalt) with missions to Poland, Albania and Moldova and with consultancies to international delegations and SMEs on the transformation to market economies. Economist and statistician at a Governmental Research Institute with focus on GDR-economy and COMECON countries. Member of International Statistical Institute (ISI), German Statistical Association, German Council of Foreign Relations, German NATO Association. Former deputy member of German Statistical Advisory Committee to the Government. Publications: Financial Crisis and Official Statistics, publ. IGI. Quality of Statistics, publ. GSMI. Economic Growth, publ. GSMI. Transformation of GDR's Economy. Contribution to the Manual of Privatization and Transformation funded by the German Ministry of Finance.

Idongesit Williams is a Post-Doctoral researcher at Center for Communication, Media and Information Technologies (CMI) located at Aalborg University Copenhagen. He holds a Bachelor in Physics, a Master degree in Information and Communications Technologies and a Ph.D. He has since 2010 researched into Knowledge Management, Organizational Learning, socioeconomic, socio-technical related to Information and Communications Technologies. He has authored and co-authored more than 30 research publications, including journal papers, books, book chapters, conference papers and magazine articles. He is the co-editor of the Books, The African Mobile Story, and the Handbook on ICT for developing countries:5G perspectives. He has organized conferences like the CMI annual conference and the CMI/GTUC conferences.

Yair Wiseman received the M.Sc. (summa cum laude) and Ph.D. degrees from Bar-Ilan University. He carried out his first Post-Doc research at the Hebrew University of Jerusalem, and the second Post-Doc research at Georgia Institute of Technology. His research interests include autonomous vehicles, intelligent transportation systems, embedded systems, real-time systems, computational transportation science, operating systems, and process scheduling. Dr. Wiseman has authored two books. He is on the editorial board of several journals, a member of dozens of conference committees and a reviewer of many scholarly journals. Dr. Wiseman has been teaching in many institutes including Bar-Ilan University, The Hebrew University of Jerusalem and Israel Aircraft Industry and has been supervising many graduate students. An interesting point is that Albert Einstein is Dr. Wiseman's academic great-great-grandfather (i.e. the advisor of the advisor of the advisor of Dr. Wiseman's advisor). Dr. Wiseman has

collaborated with other partners and received research grants to run an active laboratory from inter alia Sun Microsystems, Intel, and Polak Foundation. Dr. Wiseman is an international expert who has reviewed and evaluated several large projects of the European Union, Israel Science Foundation, Marie Skłodowska-Curie actions in Ireland and Ministry of Education and Science of Kazakhstan. His papers have been published in many venues around the world.

Sam Wong is an International Development Scholar. He obtained his PhD in Development Studies at Bradford University, UK. Before he teaches at University College Roosevelt, the Netherlands, he spent 15 years at Liverpool, Lees and Bradford Universities, UK. His research interests include development politics, power, gender and technologies. His research areas include SSA (Ghana, Nigeria, Uganda) and South Asia (Bangladesh and India).

Michelle F. Wright, Ph.D., is a Research Associate at The Pennsylvania State University and a research fellow at Masaryk University in Brno, the Czech Republic. Her research focus is on the contextual factors which influence children's, adolescents', and young adults' involvement in aggressive behaviors, with a special interest in social goals, peer status, and cultural values. She has published on these topics, with her most recent work focused on culture and anonymity, and their role in cyberbullying among adolescents.

Gen'ichi Yasuda joined Information Science and Control Engineering Course in Department of Mechanical Engineering, Faculty of Engineering, at Nagasaki Institute of Applied Science, Japan as a Research Assistant, in 1983. He was a Professor in Department of Mechanical Engineering during 1996-2005, a Professor in Department of Human and Computer Intelligence, Faculty of Informatics during 2005-2010, and also a Professor in Division of Instrument and Control, Department of Electronics and Information Technology, Graduate School of Engineering during 1996-2010. He is now a Guest Professor at Nagasaki Institute of Applied Science. He is a recipient of CACS International Automatic Control Conference Best Paper Award in 2009. He also served as session chairman in many international conferences. He originated the frontier researches on non-centralized systems and grouping robots as a doctoral dissertation at Tokyo Institute of Technology since 1971. His current research projects include a multi-microcomputer based development of automated noodle making machines.

Szde Yu is an associate professor of criminal justice at Wichita State University. He leads a Cyber Lab in conducting research on digital forensics and cyber profiling. He holds a Ph.D. in Criminology and several computer forensics certificates from the US Department of Defense as well as private sectors.

Andrej Zgank received his Ph.D. from the University of Maribor, Slovenia, in 2003 and is now an Associate Professor in the field of Telecommunications. His research interests are in the areas of advanced telecommunication services and applications design, speech and language technologies and digital signal processing. He is a member of IEEE, ISCA, and SDJT.

Tawanda Zinyama is the current Chairman and Senior Lecturer of Public Administration in the Department of Political and Administrative Studies at the University of Zimbabwe. He received a PhD degree from the University of Zimbabwe in Zimbabwe. His research interests include public administration and institutional governance, public procurement, corruption, public policy, local government

About the Contributors

management, public sector management reforms, democratic governance and public finance management. He has supervised several MPA and three PhD students in the discipline of public administration and public policy.

Index

#BlackLivesMatter (BLM) 1096-1097, 1099-1103, 1105
 'Aurah 450
 1st Dimension Divides 1682
 2nd Dimension Barriers 1682
 3D Geometric Modelling 1195
 3D modelling 1185, 1195
 3D Parametric Modelling 1195
 3rd Dimension Barriers 1682
 5G 39, 138, 265, 267, 537, 887, 914, 921-923, 933, 1440, 1450, 1928
 8×1 MUX 279, 283, 289

A

Abduction 345, 346, 351-354, 693, 1491
 Abortion 1860-1871
 Absolute Positioning 10
 Absorption 143, 419, 1056, 1255-1260, 1262-1267
 Academic Libraries 1757-1766, 1775, 1779
 Access Rate 191-200, 204, 206, 1674
 Access to Health Care 1886, 1958
 Access to ICTs 1408, 1410, 1415, 1417, 1700-1710, 1713, 1718
 Accessibility 463, 491, 559, 802-808, 1070, 1079, 1203-1204, 1209, 1354, 1355, 1361, 1398-1401, 1415, 1513-1525, 1717, 1738, 1768, 1779, 1784-1788, 1791, 1794
 Accessible Rich Internet Applications (ARIA) 1517, 1525
 Accountability 440, 446, 1099, 1135, 1158, 1357, 1359, 1409, 1471, 1487, 1490-1493, 1499, 1504-1511, 1571, 1574, 1602, 1605-1609, 1839, 1923
 ACID model 795-796
 Activation Function 14-19, 22, 26, 88, 310, 1904
 Adaptation 82, 83, 539, 547, 555, 683, 791, 796, 916, 923, 1131, 1165, 1167-1170, 1172, 1174-1175, 1177, 1398, 1401-1413, 1416-1422, 1954
 adaptive logic module 258

Adaptive Machine Translation 879, 882-883, 886
 Adaptive Streaming 606, 618
 Add-Compare-Select Unit (ACSU) 247, 248, 256
 ADS-B 673
 Advance Fee Scam (Nigeria 419) 1725
 Advocacy 190, 1230, 1463-1468, 1561, 1792, 1957
 Advocacy Journalism 1467, 1468
 ADWIN 606, 618
 affective dimensions 1199
 Affordability Test 1851
 Affordable Care Act 1850-1853, 1856-1859
 Affordable Care Act (ACA) 1850-1859
 Affordance 1650, 1661
 AFP 847, 849-855
 Aging 138, 816, 1085, 1089, 1093-1095, 1513, 1521, 1614, 1858, 1872, 1883, 1886, 1919, 1934
 Agriculture 4.0 1323-1327, 1330-1332, 1335
 Agritechnica 1327, 1329, 1332, 1333
 ai4good 766-767
 air navigation system 662, 664, 671-674
 Air Traffic Control (AC) 126, 167, 408, 656, 662, 664-665, 670, 673, 728, 733, 766, 856, 936, 1147, 1160, 1276, 1350, 1477, 1659, 1716, 1778, 1793
 Air Traffic Management (ATM) 671, 673
 AlexNet 18, 19, 23
 alternate aerodrome 668
 ALU 271-274, 287-289
 Aluminum oxide 1275, 1288
 Amazon Web Services 207, 210, 212, 216, 467
 Ambient Intelligence 151, 1931
 amplifier 986-987, 989, 994-995, 1000
 analysis of heterogeneous information 685
 And Smartphones 887, 904, 907, 1683
 Android 474-476, 479-487, 887, 892-893, 1068, 1125, 1384-1386, 1389-1391, 1395, 1396, 1661, 1781
 Android Studio 1384, 1385, 1395, 1396
 Anomaly IDS 513

Index

Anonymity 337, 341, 356, 365, 371-375, 380, 383, 385, 388, 396, 399, 409, 418, 432, 456, 904-905, 909, 1109
ANS 659, 661-662, 667-668, 670, 673, 762
Antibiotic Resistance 1366-1368, 1374-1378, 1382
Anxiety 194, 360, 364, 372, 380, 389, 401, 450, 906, 1077, 1200, 1398, 1399, 1617, 1674, 1717, 1894
Apocalypse 1638-1647
applications of NBIoT 914, 917-918, 923
Apriori 770, 773, 830
Ar 197, 199, 491, 885, 1135-1138, 1367, 1657, 1754, 1944-1950
Architecture Driven Modernization 889, 902
Architecture Driven Modernization (ADM) 888-893, 896, 898-899, 902
Architectures 19, 24, 25, 55-60, 64, 68-71, 87, 90, 94-97, 251-255, 260, 266-270, 547, 575, 577, 582, 762, 782, 813, 826, 876, 879, 881, 885, 915, 935, 939-940, 979, 988, 1000, 1014, 1049, 1258, 1266, 1902, 1905, 1930
arithmetic operations 264, 270-275, 287, 317, 319
array synthesis 953, 965
Artificial Honeybee Colony Calculation (ABC) 106
Artificial Intelligence 1, 12, 13, 24, 27, 28, 42, 54-58, 67-70, 73, 74, 87, 98, 107, 125, 127, 130, 138, 151, 152, 267, 329, 330, 574-576, 599, 601, 632, 658-659, 671-673, 682-685, 698-699, 749-751, 764-767, 777, 818, 820, 857, 868, 877-878, 881, 942, 946, 962, 1033, 1131, 1133, 1138, 1251, 1473, 1474, 1478, 1479, 1640, 1769, 1771, 1775, 1779, 1782, 1834, 1836, 1898, 1908, 1911-1915, 1918, 1928, 1931, 1939, 1940, 1944, 1950
artificial intelligent 55-60, 68, 71, 72
Artificial Neural Network (ANN) 13, 26, 1910
AS PIP 672-673
ASCII codes 274, 275, 280, 282, 289
ASIC 255-258, 265, 271, 289, 613
Assessment Methods 321, 325, 327, 1321, 1885
Assist 2, 8, 46, 193, 194, 267, 673, 907, 1126, 1128, 1131, 1322, 1354, 1403, 1429, 1482, 1571, 1672, 1678, 1689, 1759, 1770, 1775, 1842, 1949
assistance node 504-513
Association Rule Learning Algorithms 773, 777
ASSIST 673
ATC 661, 663-664, 668, 673
ATL (Atlas Transformation Language) 902
Atom 619-622, 733-734, 736-737, 739, 741-742, 744, 1259, 1270, 1366, 1428
Attack 193, 194, 344-347, 352, 411, 473-487, 490, 492, 501, 503, 514-537, 540, 821, 1366, 1408, 1499, 1508, 1558, 1791, 1954, 1959

Attacks 194, 344-347, 351, 357, 381, 401, 473-484, 487-491, 500-504, 514, 517, 519, 523-540, 825, 1153, 1412, 1491, 1508, 1731, 1891, 1901, 1927, 1938
Attitude 413, 417, 909, 912, 1073, 1132, 1389, 1454, 1491, 1492, 1510, 1608, 1717
Authentication 467, 468, 476-480, 483, 484, 491-494, 497, 498, 502, 511, 516, 521-525, 605, 1050, 1160, 1470-1474, 1773, 1895, 1896, 1899, 1922, 1931-1934
Author-me.com 1462
Automated Highway System (AHS) 10
Automated Reasoning 767, 1015
Automatic Keyword Extraction Algorithms 108, 125
Automatic Machine Learning 134, 684
automatic speech recognition 73, 74, 78, 84-86, 140, 145-152
Autoscaling 209-212, 216
AWG 980, 982-983, 985, 988-990, 994, 998, 1000

B

Back Propagation Neural Network 769
backpropagation approach 291
backstepping 42-44, 49-53
Bag of Documents Model 125
Bandwidth Delay Product (BDP) 979
Bandwidth-Delay Product 968, 970, 978
Bangladesh 215, 1342, 1343, 1351, 1413, 1419, 1481-1495
base component 551, 555-558
BBS 1462, 1468
BCD conversion 274, 275, 289
Behavior Public Policy 1537-1539, 1546
Behavioral Economics 428-435, 438, 440, 447, 1538, 1543, 1544, 1548
Behavioral Insights 431, 1538, 1539, 1545-1548
Behavioral Political Economy 445, 447
Behavioral Public Policy 1537-1541, 1545-1548
behavioural consequences 1073, 1076, 1078
Beliefs 54-56, 60-67, 71, 360, 364-373, 377, 381, 382, 385, 390, 400, 451, 460, 631, 1097, 1399, 1646, 1674, 1724
Benign tumors 847, 855
BER 245, 991, 995, 1000, 1053
BHW 1352

Big Data 214, 216, 411, 417, 419, 430-438, 447-452, 455, 460, 471, 491, 499, 536, 539, 559-562, 571-589, 592-602, 616, 617, 670, 731-732, 744-746, 749-750, 768-770, 775-777, 786-788, 791-794, 796-799, 823-824, 826, 869, 899, 929, 1062, 1141, 1147-1149, 1207, 1452, 1478, 1479, 1782, 1789, 1794, 1795, 1813, 1836, 1901, 1909, 1914, 1915, 1918, 1924-1931, 1934, 1940, 1941, 1950

Big Data Analytics 491, 536, 574, 575, 578, 583-589, 592-601, 617, 746, 768-769, 777, 786, 929, 1478, 1813, 1914, 1929, 1941

Big Data Mining 581-584, 768, 775-776, 824, 826, 1930

big data problems 577, 585, 598

BIM Modelling 1182, 1195

binge watching 1071, 1075-1077, 1082, 1084, 1091

Bio Energy 1326, 1335

Bio Mass 1335

Bio-based innovations 1323, 1325, 1331, 1332

Bioeconomy 1323-1327, 1330-1335

biogeography-based optimization 950-951, 960-965

Biomass 656, 1290, 1306, 1321, 1323, 1326-1332, 1368

bi-partite graph 828-830, 843-844, 846

BIRCH 563-565, 575, 576, 883

bits'-carrying capacity 967-970, 972-974, 976, 979

Black Box 90, 177, 552, 571, 632, 1903-1906, 1911, 1915, 1918

black phosphorus (BP) 1255, 1263, 1270

BLM 1096-1097, 1099-1102, 1105

Block RAM 264, 270

Block RAM (BRAM) 264, 270

Blockchain 164, 173, 176, 569, 574, 813, 1456, 1459

Blogs 390, 398, 399, 433, 745, 884, 1099, 1153, 1156, 1159, 1409, 1422, 1687, 1762, 1768, 1793

Body-Area Network 1943

Boko haram 344-347, 351-355

Bokoharamism 346, 347, 354

Bomb 344, 347, 354, 355, 1426, 1428, 1433, 1434, 1791

bonding social capital 1111-1113, 1115

Bottom Days 1639, 1646

branch metric unit 247-250, 256

Branch Metric Unit (BMU) 247, 256

bridging social capital 1111-1114

Bring Back Our Girls 345

Broadcast 84, 348, 502, 505, 507, 514-520, 523, 527, 533-536, 601, 888, 974, 1007, 1009, 1022, 1030-1031, 1036, 1053, 1056, 1070, 1081, 1091, 1130, 1132, 1159, 1461, 1464, 1465

Bulk Density 1343, 1348

bully 356, 364, 375, 388, 389, 395, 397, 400, 403

Bully Victim 389, 397

Bullying 333, 344, 356-408, 427, 1449

BVLOS 673

C

CAP Theorem 792, 796

Capitalism 449, 1552, 1563, 1574, 1638, 1643-1647, 1692, 1782, 1789, 1790, 1794, 1795

Carbon Resin Electrodes 1250, 1363, 1379

Casuistry 1453, 1459

causal models 625, 631-635

Causality 751, 1906

CBS 1087, 1093, 1725

Centroid 565, 758, 830, 1751, 1756

Change Management 547, 557, 1169-1170, 1404, 1447

channel assignment 1018-1019, 1024-1035

Chemical Oxygen Demand 1383

Choice Architecture 1537, 1548

chronemics 905, 908-909, 911

Chronic Problem 411, 412, 424, 426

Citizen 348, 1087, 1090, 1132, 1146, 1152-1156, 1159-1162, 1402, 1412, 1464, 1481, 1492, 1495, 1497, 1500, 1502, 1508, 1520, 1525, 1538, 1550-1555, 1563, 1564, 1570, 1571, 1574, 1579, 1582-1597, 1605, 1609, 1622, 1677, 1831, 1835

Citizen Journalism 1152-1154, 1159-1160, 1162, 1677

Class I-III 1826

CLI 210, 1118, 1129, 1418

Client 51, 217-228, 232, 464, 467, 480, 495-498, 515, 553, 562, 572, 595, 601, 602, 613, 618, 750, 980, 982-983, 1019, 1024-1025, 1037-1038, 1043, 1050, 1388, 1391, 1396, 1405, 1669, 1738, 1768, 1925

Client-Server Architecture 1396

Climate Change 1133, 1325, 1329, 1353, 1398-1421, 1554, 1562, 1962

Climate Finance 1401-1406, 1420-1422

Clinical 395, 401, 407, 443, 701, 847-848, 853, 857, 939, 1040, 1079-1080, 1083, 1092, 1340, 1348-1351, 1367, 1374, 1377, 1380, 1381, 1468, 1474, 1726, 1797-1816, 1820, 1823-1827, 1833, 1834, 1860, 1862, 1884, 1885, 1888, 1889, 1892-1903, 1906-1918, 1923, 1930, 1936, 1939, 1952, 1955, 1956, 1959, 1960, 1963

Clinical practices 1812

Clinical Research 1474, 1797-1805, 1810-1814, 1825, 1888, 1889, 1916, 1959

Clinical Trial 1797-1803, 1809-1814

Index

- Cloud Computing 31, 39-41, 98, 105, 106, 181, 190, 193, 194, 204-208, 214, 215, 218, 224, 231, 242, 411, 415, 419, 462-472, 578-581, 585, 597, 600, 601, 680, 767, 787, 887, 892, 900, 928-930, 938, 1048, 1769-1772, 1775, 1781, 1848, 1920, 1925-1931, 1946, 1950
- cloud evidence 467
- cloud forensics 465-472
- cloud service consumer 195, 199, 204, 206
- cloud service provider 194, 195, 206, 464, 466
- Cloud Services 32, 40, 190-194, 197-201, 204-206, 464, 467, 1925-1928, 1936, 1940
- cloudlet 28-41
- Cluster 110, 112, 121, 233, 241, 242, 260, 267, 272, 500-511, 515, 518-521, 561-573, 578, 605, 755-756, 761, 767, 793-797, 799, 828-832, 834, 836-837, 840, 842-846, 1019, 1034, 1038, 1040, 1042-1043, 1046-1048, 1050, 1300, 1303-1304, 1368, 1417, 1580, 1836, 1881-1883, 1913
- Cluster-based routing 509-513
- Clustering 12, 99, 105, 121, 123, 237-243, 502, 515, 559-576, 605, 685, 696-699, 753-756, 758-761, 764-765, 767-770, 772, 776-777, 782, 788, 818, 828-837, 839, 842-846, 870, 1038, 1040, 1047, 1464, 1473, 1535, 1917
- Clustering Algorithm 121, 502, 559, 563-576, 759, 772, 777, 828, 830
- CNS 662, 671, 673-674
- C-n-tuple 687, 689-690, 694-695, 699-700
- CoAP 926-927, 933-937
- Coarse-Grained Granularity 270
- Coase Theorem 1618, 1622
- Cognitive Architecture 55, 57, 60, 69, 71, 1200
- Cognitive Biases 1471, 1540, 1546, 1548
- cognitive model 54-57, 60, 63-71
- Collaboration 27, 28, 39, 46, 99, 178, 182, 218, 224, 346, 392, 403, 577, 582, 595, 600, 658, 670, 793, 881, 910, 919, 1097, 1108, 1110, 1200, 1245, 1326, 1337, 1346, 1347, 1378, 1412, 1479, 1506, 1529, 1558-1561, 1570, 1574, 1578, 1579, 1585, 1586, 1592, 1593, 1665, 1768, 1769, 1779, 1817, 1888, 1889, 1892, 1952, 1958
- Collaborative Decision Making (CDM) 660, 673
- Collaborative Filtering 191, 192, 197, 206, 676, 682-684, 1115
- Collective Action 1096-1097, 1101-1102, 1436, 1437, 1447-1450
- Collective Behavior 1100, 1437
- Collectivism 372, 432
- Collisions 4, 155, 531, 692-693, 700, 1053, 1279
- Command and Control (C2) 673
- Comments 333, 388, 1109, 1155-1157, 1162, 1424, 1462, 1464, 1587
- commit protocol 800-802, 804, 810-812
- common random numbers 167, 168, 176
- Commune 1589, 1591
- communication engineering 939-940, 942-944, 946, 948-949
- Communication Link 233, 234, 970, 1002, 1026, 1059
- Communication Network 233, 243, 244, 537, 817-818, 821, 824-825, 946, 970, 979-980, 1003, 1005, 1007, 1016, 1773
- Communication Networks 123, 124, 138, 471, 524, 812, 816, 820, 913-914, 921, 936, 939-945, 948, 969, 979, 1000, 1002, 1465, 1741, 1745, 1920
- Communication Time 243
- Communication, Navigation, and Surveillance (CNS) 674
- community informatics 1440, 1680
- Community Mobilization 1437, 1440, 1447, 1450
- Comparative Analyses 1883
- Compensations 1612, 1614, 1617
- Competence 400, 848, 1333, 1551, 1643, 1658, 1666, 1685, 1731, 1732, 1739, 1740, 1743, 1744
- Competent Authority 1816, 1820, 1824, 1826
- Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) 1157, 1163
- Complexities 577, 939-942, 946, 949, 1846
- Compliance 37, 218, 219, 224-232, 430, 435, 627, 1488, 1506, 1521, 1522, 1602, 1607, 1783, 1788, 1816, 1894, 1923
- composite materials 1254, 1259, 1262-1266
- Compress 207, 210, 212, 216, 468, 782
- Computational Model 54-56, 71, 158, 1285, 1307
- Computational Social Science 584, 1476
- computer forensics 337, 338, 341, 342, 467
- Computer Implementation 163, 685
- Computer Security 470, 539, 1472, 1847
- Computer Simulation 155-164, 173-176, 1250, 1304, 1307
- Computer Vision 11, 12, 24, 25, 68, 96, 128-130, 133-137, 1207, 1634
- Computer-Mediated Communication 459, 904, 908-912, 1114-1117, 1466, 1679, 1680
- concept drift 676-678, 682, 684
- concept mining 856-858, 868, 871
- Concept of Operations (CONOPS) 1556, 1574
- Concourse 1746-1749, 1752, 1756
- concurrency control 801, 804, 809, 811-814
- condenser 1272-1280, 1288
- Conditions of Instruction 1750, 1756

- configurable logic block 258, 264, 270
 - Configurable Logic Block (CLB) 270
 - Conformance 218, 219, 224-232, 891
 - Confusion Matrix 113, 114, 851-852, 855
 - Consent 399, 409, 421, 1441, 1454, 1460, 1482, 1808-1811, 1839-1843, 1848, 1864-1867, 1871, 1895, 1898, 1899, 1939
 - Consequence 48, 145, 146, 265, 328, 363, 411, 419, 421, 426, 437, 453, 581, 691-693, 695, 739, 800, 804, 814, 878, 880-881, 1107, 1113, 1203, 1225, 1272, 1278-1279, 1372, 1374, 1624, 1626, 1632, 1678, 1687, 1703, 1748, 1863
 - Consistency 395, 419, 420, 426, 446, 516, 549, 584, 628, 692, 700, 790-791, 794-797, 799-801, 804, 813, 1473, 1707
 - Consumer Protection 1538-1540, 1543-1548
 - controller design 6, 43
 - Convenience 415, 455-458, 475, 481, 724, 904, 1052, 1069-1071, 1075, 1079, 1081, 1084, 1175, 1622, 1656, 1659, 1751
 - Conversational journalism 1135, 1138
 - Convolution Neural Network 774
 - convolutional codes 246, 253-256
 - Convolutional Encoder 245, 256
 - Convolutional Layer 16-19, 26, 1904
 - Convolutional Neural Network 12, 13, 16, 17, 24-26, 134, 152, 826, 1907, 1912, 1914, 1917
 - Convolutional Neural Network (CNN) 12, 26, 134
 - Convolutional Neural Networks 12, 13, 16, 18, 22-25, 97, 146, 151, 762, 1911, 1914, 1917
 - Cord-Cutting 1095
 - Corpus-Based Machine Translation 877, 882, 886
 - Correspondence Analysis 328, 1874, 1880, 1881, 1886
 - Corruption 416, 517, 1134, 1482, 1489-1491, 1494, 1497-1511, 1585, 1602, 1639, 1838
 - Coupling 217-232, 545, 930, 987, 1049, 1259
 - CPU 30, 31, 46, 98, 156, 209, 211, 266, 272, 289, 479, 569, 570, 602, 603, 607, 608, 613, 616, 618, 843-844, 932, 1039-1041, 1043-1045, 1047, 1937
 - Crispr 1954, 1960
 - Critical Information Literacy 1682
 - Crossfire 1481-1487, 1490-1494
 - Cross-Platform Development 473, 887, 893, 896
 - crowdsourcing approach 1199, 1207, 1209
 - Crypto-Currencies 1456, 1459
 - C-system 687-688, 690-692, 695-696, 699-700
 - CT scan 847, 849, 855
 - CUDA 608, 609, 614-618
 - cudaMemcpyToSymbol 608, 618
 - Cultural Diversity 1398-1401, 1405, 1406
 - Cultural Heritage 107, 115, 122, 129, 130, 133-137, 437, 1181, 1194, 1447
 - Culture 76, 115, 127-132, 135, 177, 179, 366, 390, 394, 398, 404, 415, 416, 419, 451, 456, 580, 730, 1077-1078, 1096, 1102, 1104, 1156, 1158-1161, 1343, 1353, 1354, 1398-1401, 1404-1406, 1409, 1416-1419, 1433, 1447, 1452, 1458, 1476, 1477, 1492, 1503, 1529, 1572, 1597, 1599, 1608, 1635, 1638, 1639, 1644, 1645, 1648, 1650, 1658-1661, 1669, 1676, 1677, 1680, 1682, 1695, 1704-1706, 1713, 1741, 1745, 1782, 1787, 1792-1794, 1868, 1957
 - Customization 147, 148, 1132, 1889
 - cyber context 342, 356, 357, 361, 365, 381
 - cyber profiling 333-342
 - cyberbullying 339, 356-409, 418
 - Cyberization 344, 347, 355
 - Cyberspace 334, 339, 341, 345, 348, 351, 354, 355, 369, 370, 381, 384, 392, 396, 415, 582, 1080, 1150, 1467, 1680, 1768, 1775
 - Cyberterrorism 344
- ## D
- Dapchi girls 353
 - Data Analysis 133, 176, 207, 313, 562, 573-584, 593, 596-598, 603, 605, 609, 624, 625, 634, 636, 637, 639, 655-656, 727, 750, 768, 775, 782, 789-790, 798, 815-816, 825, 1134, 1141, 1143, 1151, 1477, 1545, 1620, 1748, 1751, 1810, 1885, 1908, 1915, 1923-1926, 1930
 - Data Breach 458, 460
 - Data Cleaning 818, 1142-1143, 1150
 - Data Cleaning or Data Scrubbing 1150
 - Data Compilation 1142-1143
 - data conflicts 800, 804, 808-809
 - data journalism 1134, 1138, 1140-1150, 1162
 - Data Mining 125, 547, 559, 562, 573, 574, 581-584, 615, 661, 663, 670, 675, 681-684, 745-746, 749-755, 763, 766, 768-769, 775-778, 782, 784, 791, 800, 815-816, 818-826, 828, 844-851, 853-854, 856, 868, 870, 872, 1229, 1251, 1477-1480, 1657, 1813, 1909, 1920, 1923, 1924, 1928-1931
 - Data Protection 411-413, 416, 417, 424-428, 449, 462, 468, 473, 493, 500, 514, 528, 588, 763, 1790, 1831, 1834-1838, 1846
 - Data Science 573, 596, 599, 663, 670, 699, 812, 1473, 1657, 1897-1900, 1919, 1940
 - Data Scraping 1142, 1150
 - Data Stream 573, 581, 587, 602, 605-618, 675-677, 679-682, 823, 825, 1122

- Data Stream Management System 615, 618
- Data Stream Management System (DSMS) 618
- Data Streams 581, 584, 604-610, 614-618, 675-684, 957, 1895, 1925
- Data Understanding 1142-1143
- Data Validation 797, 1142-1143, 1150
- Data Visualization 1142-1144, 1149-1151, 1827, 1836, 1925
- Database 18, 22, 69, 97, 115, 125, 144, 182-187, 192, 229, 453, 455, 458-461, 477, 484, 487, 495, 500, 503, 508, 509, 544, 553, 557, 560, 574, 584, 587, 588, 615-618, 642-643, 662-663, 667, 670, 674, 685, 697-699, 776, 778-791, 793-800, 802, 804, 807, 809-815, 818, 843, 847, 857, 868, 870, 878, 886, 1040-1041, 1048, 1195, 1217-1218, 1222, 1225, 1228-1229, 1323, 1331, 1384-1390, 1393-1397, 1412, 1452, 1459, 1462, 1472, 1530, 1531, 1590, 1620, 1686, 1725, 1758, 1759, 1763, 1773, 1776, 1789, 1824, 1826, 1831, 1841, 1843, 1890, 1891, 1896, 1901, 1924
- Dataset 114, 118-121, 125, 130, 199-203, 218, 563-567, 570, 571, 574, 578, 595, 600, 637, 642, 759, 831, 840-841, 844, 848, 850-855, 1142-1144, 1150-1151, 1874
- Datasets 107, 108, 113-118, 121, 122, 126, 135, 411, 563, 566, 567, 571, 572, 576, 583, 593, 639, 655, 775, 783, 840-841, 844, 847-848, 853, 1141, 1143, 1145, 1772, 1895, 1896
- Dbms 92, 618, 786, 790
- DBSCAN 568-570, 576
- DDP 1595
- Deadlock 48, 53, 418, 1004-1005, 1013
- Death in Custody 1487, 1494
- Decentralisation 1506, 1507, 1597
- Decentralization 1587, 1588, 1597-1611, 1873, 1886
- Decision Making 10, 59, 60, 193, 194, 310, 312, 319, 419, 423, 424, 428-433, 438-441, 444-447, 501, 515, 542, 580, 587, 658-661, 663, 668-674, 752, 776, 815-816, 818, 909-910, 930, 962, 1309, 1351, 1399, 1403, 1557, 1561, 1570, 1584, 1701, 1702, 1712, 1718, 1733, 1737, 1742, 1826, 1838, 1850, 1872, 1901, 1903, 1907, 1908, 1912, 1916, 1918
- Decision Support System 557, 659, 674, 1894, 1899
- Decision Support System (DSS) 659, 674
- Decision Tree algorithm 771, 777, 851
- Deep Boltzmann Machine 90, 91, 97, 774
- Deep Learning 12, 24-26, 87, 90, 95, 97, 121, 130, 133, 146, 152, 270, 569, 571, 574, 599, 613, 749, 762, 764, 773, 777, 1472, 1473, 1477, 1635, 1657, 1901-1905, 1909-1916, 1945, 1950
- Deep Learning (DL) 12, 26
- Deep Learning Algorithms 571, 773, 777, 1909
- Deep Neural Network 14, 24, 26, 134, 1904, 1912
- Deep Neural Network (DNN) 26
- Defence 354, 415, 490, 526, 1411, 1429, 1430
- Definitional Context 858, 872
- defuzzification 317, 318, 322, 329-331
- Dementia 1089, 1092-1093, 1095, 1197, 1911
- Democracy 1153, 1417, 1466-1469, 1474-1476, 1492, 1494, 1559, 1571, 1574, 1578-1590, 1593-1595, 1599-1601, 1611, 1620, 1643, 1645, 1672
- Denmark 441, 1068, 1342, 1436, 1440, 1514, 1517, 1518, 1538, 1820
- denotational mathematics 54-57, 60, 67, 70, 72
- DenseNet 21
- Deoxyribonucleic Acid 730, 747, 1938
- Depression 360, 361, 364, 369, 380, 383, 389, 395, 401, 407, 409, 1073-1074, 1076, 1089-1090, 1093, 1302, 1641, 1642, 1828, 1829, 1832, 1834, 1837, 1849, 1855
- Descriptive Analysis 625, 628, 633, 634, 643
- Design Process 23, 53, 56, 1177, 1197-1198
- Design Recovery: Creating design abstractions from a combination of source code 1230
- Designer Babies 1958
- Deterministic Algorithm 164, 176
- Deterministic Sequence 164, 176
- Developing Countries 181, 913-915, 918-919, 921, 1067, 1325, 1330, 1332, 1337, 1340-1351, 1362, 1408-1410, 1414, 1415, 1418-1421, 1441, 1442, 1581, 1674, 1712, 1715-1721, 1725, 1898, 1900
- Devolution 1597-1611
- Diabetes 1088, 1689, 1888-1900, 1938
- Diagnose 722, 853, 1072, 1204, 1797, 1819, 1905, 1906, 1915
- Diagnostic 664, 725, 847, 855, 1205, 1438, 1446, 1726, 1805, 1812, 1818-1820, 1824, 1825, 1873, 1882-1885, 1891, 1896, 1901, 1906-1909, 1914-1918, 1937
- Digital Age 409, 430-438, 909, 1083, 1131, 1146, 1159, 1196, 1418, 1466, 1467, 1479, 1676, 1680, 1684, 1715, 1757-1759, 1763-1766, 1771, 1779
- Digital Audio Broadcasting (DAB) 957, 966
- digital consumption 1066-1067, 1072, 1078-1080, 1083-1084
- Digital Corpus 1783, 1788, 1795
- Digital Divide 1087, 1102, 1524, 1660, 1672-1684, 1689-1700, 1703, 1704, 1714, 1717, 1718
- digital document 1118, 1129
- digital footprints 333-343
- Digital Forensics 341, 342, 462-472, 1696
- Digital Humanities (DH) 1795

Digital Immigrants 912, 1660
 Digital Literacy 364, 1648, 1649, 1654, 1658-1661, 1678, 1681-1698, 1717
 Digital Media 348, 365, 574, 1066, 1068, 1079, 1081, 1085, 1087, 1096, 1106, 1118, 1130-1132, 1135-1138, 1140, 1152, 1196, 1466, 1467, 1582-1585, 1672-1678, 1682, 1684, 1694
 Digital Natives 912, 1649, 1650, 1656-1660, 1679, 1697, 1699
 Digital Participation 1593, 1682
 Digital Signal Processing 149-152, 265, 266, 270
 Digital Signal Processing Block (DSP) 270
 Digital Social Network 456, 1109, 1117
 Digital Society 431, 1516, 1686-1688, 1693, 1696, 1698
 Digital Technologies 128, 132, 356, 357, 363, 365, 412, 887, 1083, 1085, 1134, 1196, 1323, 1325, 1328-1332, 1335, 1449, 1583, 1649, 1650, 1658, 1757, 1758, 1761, 1889
 Digital Technology 356, 360, 361, 364, 365, 371, 385, 388, 816, 1083, 1086, 1091, 1130-1131, 1208-1209, 1467, 1660, 1683, 1686, 1687, 1691, 1693, 1698, 1746, 1757
 Digital Video Broadcasting (DVB) 957, 966
 Digitization 457, 1095, 1106, 1134, 1323, 1327-1332, 1760, 1765, 1766, 1779, 1784-1786, 1789, 1790, 1793, 1794, 1888
 Dimension of the Space 747
 Direct Democracy 1581, 1595
 Directional Motion of the Fluidized Bed 1289
 Directive 824, 1516, 1523, 1524, 1538, 1547, 1816-1820, 1823-1826, 1837, 1961
 Disaster 352, 593, 600, 816-817, 912, 919-920, 1409, 1412, 1413, 1416-1422, 1490, 1612-1622, 1625, 1630, 1632, 1638, 1641, 1644, 1793
 Disaster Management 919-920, 1409, 1412, 1413, 1416, 1419, 1422, 1490
 Distributed Computing System 241, 244
 Distributed Decision Support System (DDSS) 674
 Diversity 21, 75, 223, 261, 463, 891, 930, 934, 951, 1023-1024, 1099, 1116, 1132, 1150, 1155, 1368, 1398-1401, 1405-1407, 1534, 1561, 1599, 1618, 1656, 1740, 1782-1784, 1788-1792, 1841, 1927
 DM 659-664, 668, 670, 674, 755, 1311-1313, 1320
 D-n-tuple 687, 690, 696, 700
 Documents-Words frequency matrix 832, 839, 842-843, 846
 doing gender 388-391, 396, 397
 doxing 398-401
 DPP 1579, 1585-1587, 1593

Drawing 127, 131, 132, 135, 347, 348, 354, 374, 387-391, 430, 435, 786, 855, 975, 1100-1101, 1181, 1203, 1388-1393, 1408, 1411, 1413, 1655, 1677, 1701, 1938
 Drone 207-216, 674, 1130, 1132-1133, 1138, 1140, 1147
 Drone journalism 1130, 1132-1133, 1138, 1140, 1147
 DRTDBS 800-802, 804, 807, 809, 812, 814
 Drugs 849, 1366, 1380, 1454, 1488, 1492, 1602, 1797-1800, 1812, 1824, 1867, 1896, 1939
 Drying 1290, 1297, 1301-1305, 1307, 1346
 DSN adoption 1106, 1114
 D-system 687-688, 690-691, 700
 DVR technology 1095
 Dynamic Load Balancing 98, 100, 106, 240
 Dynamic Model 60, 1322, 1913

E

Economic Crisis 1662, 1663, 1666, 1671
 Economic Development 1113, 1117, 1405, 1504, 1505, 1534, 1597, 1598, 1604, 1611, 1617, 1662-1670, 1704, 1718, 1719
 Economic Recovery 1671
 Economy 179, 319, 439, 445-447, 450, 596, 602, 625, 831, 1077, 1099, 1114, 1309, 1321, 1325, 1326, 1331, 1354, 1387, 1418, 1421, 1447, 1459, 1497, 1504, 1526, 1527, 1532-1537, 1551-1555, 1562, 1563, 1572, 1575, 1578, 1598, 1619, 1662-1671, 1686, 1689, 1690, 1701, 1703, 1712, 1715, 1724, 1732-1734, 1738, 1822, 1855, 1857, 1928, 1940
 Ecore metamodel 891, 893, 902
 Ecosystem Services 1421, 1551, 1572, 1574
 E-Democracy 1579, 1582-1585, 1593-1595
 Edge Computing 41, 266, 267, 1927-1930
 e-governance 1408, 1420, 1524, 1593, 1595, 1696
 E-Government 1440, 1450, 1513-1525, 1583, 1585, 1592-1595, 1689, 1693, 1697, 1698, 1702, 1715
 EJB Component 553, 557
 Elasticity 98, 463, 472, 581, 932, 1562
 Elderly 27, 138, 147, 527, 595, 1085-1095, 1197, 1211, 1409, 1413, 1521, 1748, 1882
 Electrochemical Treatment Technique 1363, 1383
 Electrocoagulation 1363, 1372, 1373, 1383
 Electrodeposition 1363, 1383
 Electroflotation 1363, 1373, 1383
 Electrolysing Equipment 1363
 Electromagnetic Interference (EMI) Shielding 1270
 Electronic medical record 1888, 1899
 Electronic Resources 1757-1760, 1766
 Electrophoretic Deposition 1383

Index

Element at Risk 1624, 1632, 1636
Embarrassingly Parallel 40, 613, 616, 618
Embedded System 141, 1775
Emerging Pollutants (EPs) 1365, 1383
Emmanuelle Charpentier 1952
Emotion 319, 337, 397, 538, 906, 1078, 1094, 1198, 1205-1206, 1211-1212, 1469
Empathy 360, 364, 373, 378, 396, 1100, 1204, 1732
encryption 215, 262, 266, 416, 480, 487, 491, 498, 517
Endocrine 1369, 1375, 1378, 1888, 1889, 1900
end-to-end delay 970-971, 979, 1026, 1029, 1031
energy conservation 534
Energy Consumption 30-32, 38, 253, 266, 267, 501, 508-510, 521-527, 533-535, 538, 631, 820, 826, 944-945, 1005, 1056, 1060, 1187, 1244, 1327, 1329, 1372, 1373, 1670
Energy efficient IoT 923
Enforced Disappearance 1488, 1489, 1494
Ensemble Algorithms 575, 774, 777
E-Participation 1516, 1583-1586, 1593-1595
Epi 1137, 1344, 1352
Epidemiology 635, 1093, 1827, 1836, 1837, 1885, 1916
Estonia 1342, 1514, 1517, 1520, 1521
Ethical Analysis 6, 417, 418, 421, 426, 1619
Ethical Matrix 411, 417-423, 426
ethical risk 415, 427
Ethical Standards 1726
Ethical Technology 1955, 1966
Ethics 411-418, 421, 425-427, 431, 459, 762, 1150, 1424, 1425, 1429, 1432-1436, 1451-1454, 1457-1461, 1466, 1469, 1477, 1481, 1497, 1498, 1501, 1502, 1510, 1511, 1639, 1680, 1686, 1694, 1780, 1797, 1801, 1804, 1808-1810, 1813, 1815, 1827, 1839, 1840, 1846, 1849, 1850, 1860-1872, 1898, 1934, 1936, 1952, 1956, 1957, 1964, 1965
Ethics of Norms 1453, 1460
Ethics-Based Remedy 417, 426
EUDAMED 1816, 1824, 1826
European Commission 1323, 1324, 1513, 1531, 1534, 1538, 1539, 1815, 1816, 1824
European Union 448, 1323, 1329, 1332, 1334, 1408, 1418, 1513-1516, 1522, 1529-1531, 1534, 1535, 1538, 1547, 1694, 1790, 1815, 1823-1826, 1835, 1837
Europeana 115, 123, 126, 1782, 1786, 1788, 1792
Euthanasia 1861-1871
Evacuation 538, 1412, 1413, 1416, 1613, 1619, 1622, 1864
evaporator 1272-1282, 1285, 1288
Evidence-Based Policy 1548

Evolutionary Algorithms 950-951, 953, 958, 962, 1034, 1905
Evolutionary Computing 597, 601
Evolutionary Learning 1904
Execution Time 29-41, 244, 248, 565, 570, 594, 805, 1019
Existential Philosophy 1460
Exploit 28-31, 411, 451, 474, 478, 479, 484, 489, 499, 583, 613, 678, 685, 792, 799, 888, 893, 1145, 1157, 1265, 1529, 1650, 1657, 1668, 1670, 1733, 1772, 1926
Exploits 415, 474, 477, 482-484, 491, 695, 1051, 1926
externalizing difficulties 360, 373, 380
Extrajudicial Killing 1482, 1487, 1490-1494
extreme-value distribution 704-705, 707, 709, 712, 714, 716, 726-727, 729

F

Facebook 22, 334, 335, 339-342, 358, 369, 376, 399, 408, 416, 433, 449-460, 464, 467, 493, 495, 556, 560, 778, 792, 880, 906, 908, 912, 939, 1067, 1072, 1080, 1096, 1098-1099, 1101-1106, 1109-1110, 1112-1113, 1115, 1117, 1132, 1152, 1458, 1462-1470, 1473, 1476, 1579, 1585, 1591, 1638, 1651, 1661, 1679, 1681, 1697, 1762, 1768, 1782, 1829-1831, 1837, 1841, 1844
Factor Analysis 625, 630-634, 663, 1746, 1747, 1751, 1752, 1756, 1880
Factory Analysis 636
Fair Use 1788, 1792, 1795
Fake News 406, 430, 1467-1480, 1678
False Negative 340, 342
false positive 340, 343, 503
Far Infrared Sensors (FIRS) 11
Farḍhu kifāyah 1867, 1870
Fatāwā 1860-1863, 1866-1870
Feature Extraction 128, 140, 143-145, 149-153, 723, 844-845
feature map 16, 17, 22, 26
Federated Search 1769, 1773-1780
Feminization 344, 346, 355
fiber delay lines 988, 997, 1000
Fiber-Wireless Enhanced Networks 41
field programmable gate array 267, 289
Field-Programmable Gate Array (FPGA) 270
Fine-grained granularity 270
Firm Distributed Real-Time Transaction (Firm DRTT) 814
fixed effects 638, 640-641, 646-648, 652-655, 657
Flagged on a Factor 1756

Flash 262, 891, 1125, 1413, 1519, 1522, 1525
 Flight and Flow Information for a Collaborative Environment (FF-ICE) 674
 flight emergency 664-665, 670
 Fluidised Bed 1305, 1307
 FNAC 853, 855
 Fog Computing 928-929, 1927, 1944-1951
 Follower 21, 1098, 1102, 1105, 1110
 Forest Ecological Management 1311, 1321
 Formal Concept Analysis 760-761, 764, 1572
 formal method 967
 Formal Methods 42, 888-889, 893, 899, 1002-1004, 1014-1016
 Formal Verification 1002, 1004-1005, 1011, 1013-1016
 forward engineering 888-890, 896, 1216, 1230
 Forward Engineering: Moving from the high-level abstractions and logical implementation-independent designs to the low-level physical implementation of a system. 1230
 FPGA 6, 254-275, 282, 283, 286, 289, 617
 FPGA routing 260, 268
 Fractional Factorial 1363, 1364, 1371
 Freedom Of Expression 401-404, 409, 449, 458, 1461, 1795
 Friction Factor Formulae 1231, 1243, 1250-1252
 fully connected layer 17, 18, 22, 26
 Future Outcome 703, 729
 Fuzzy Equations 291, 302, 309-313, 1284, 1288
 Fuzzy Logic (FL) 319, 330
 fuzzy matrix equation 291
 Fuzzy Nonlinear Systems 311, 312, 1284, 1286
 Fuzzy Number (FN) 316, 330
 Fuzzy Set (FS) 319, 330
 Fuzzy Sets 295, 296, 310-312, 316, 329, 330, 582

G

Gallica 1782, 1786, 1792
 Gatekeeper 1461, 1465-1468
 gatekeeping 1152, 1154-1155, 1159, 1161-1163
 GDBMS 778, 780-785
 Gender 81, 82, 108, 140, 144-152, 334, 344, 345, 350, 354, 360, 372, 377, 382, 384, 387-409, 437, 460, 626, 628, 641, 850, 910, 1101, 1106, 1398-1422, 1469, 1578, 1674, 1679-1682, 1699-1708, 1713-1718, 1725, 1750, 1828, 1832, 1833, 1842, 1851, 1852, 1863, 1866, 1889-1891
 Gender Difference 1699, 1700, 1703-1706, 1713, 1714, 1718
 Gender Differences 360, 377, 387, 390-394, 406, 1420, 1699, 1704, 1713, 1716, 1717, 1851
 Gender display 390, 394, 397
 Gender Equality 400-405, 409, 1401-1406, 1416, 1420, 1422, 1578, 1714
 Gender Gap 408, 1700-1704, 1716, 1852
 Gender Inclusion 1403, 1407, 1417, 1705
 Gender Mainstreaming 1403, 1407, 1414, 1417, 1422
 Gene Editing 1952, 1953, 1957, 1958, 1961-1964
 General adversarial network 97
 generalized operations and relations 689-690, 698, 700
 Generalized Theory of Relations 685
 Genetic Algorithm 158, 237-244, 503, 937, 1023-1024, 1026-1029, 1031, 1035, 1473, 1913, 1914
 Genetic Algorithms 12, 158, 162, 242, 951, 966, 1028, 1477, 1904, 1914, 1918
 Genome Editing 1952-1959, 1963-1966
 Genomics 1888, 1895-1900, 1909, 1963, 1964
 Geographical Databases 1384, 1388, 1395, 1396
 geo-informatics 1206
 Germline Editing 1956-1959, 1966
 GIS 185, 528, 538, 561, 575, 1183, 1195, 1205-1207, 1209-1210, 1330, 1411, 1412, 1624, 1627, 1630-1636
 Global Navigation Satellite Systems (GNSS) 674
 Google Books 1784-1794
 GoogleNet 19, 20
 GPU 146, 209, 255, 569-576, 603-610, 613, 616, 618
 Grade Point Average (GPA) 315, 330
 Granularity 47, 111, 258, 263, 266, 270, 463, 982-983
 granulation 1289-1290, 1297, 1299-1300, 1302-1305, 1307
 Graph 14, 48, 97, 110, 111, 125, 222, 223, 229, 235, 236, 243, 256, 316-318, 330, 331, 412, 418, 433, 434, 514, 563, 568, 569, 575, 609, 672, 677, 689, 778-790, 794, 828-831, 843-846, 867, 940-941, 947, 978, 995, 1023, 1030, 1195, 1209, 1219-1223, 1473, 1795, 1828, 1838, 1880, 1925
 graph data partitioning 782-783, 789
 Graph Data Warehouse 787, 789
 graph database model 779, 781, 783-784, 790
 graph databases 778, 780-789
 Graph Mining 609, 784, 790, 828, 830-831, 844-846
 Graph Pattern Matching 782, 790
 Graph Query Processing 783, 790
 Graph Theory 672, 778, 940-941, 978, 1828, 1838
 graphene 1254-1255, 1258-1260, 1263-1270
 Graphics Processing Unit (GPU) 569, 603, 607, 609, 616, 618
 Green Climate Fund 1403-1407
 green IoT 920, 923
 Grey Number (GN) 330
 Grey Numbers 315, 316, 319, 329

Grey System (GS) 319, 330
 Grey Systems 329
 Group Knowledge Model 1574
 GUI 884, 1126, 1129, 1668

H

Hand Rotation 1756
 Hard Distributed Real-Time Transaction (Hard DRTT) 814
 hash code 505-507, 520-524, 527, 529, 532-536, 540
 Hash Code Generation 527
 hash table 609, 616, 618
 Hash Table (Hash Map) 618
 HathiTrust 1760, 1782-1794
 Haxe 887, 889, 891, 893, 895-902
 Haxe language 891, 896, 901
 hazard 1339, 1380, 1404, 1412, 1623, 1624, 1630-1636, 1816
 HCC 847-851, 853
 Health Care 393, 577, 588, 595, 596, 847, 1336, 1337, 1347, 1350, 1352, 1719, 1797, 1831, 1832, 1835-1839, 1845, 1846, 1853, 1859, 1872-1874, 1882, 1883, 1886-1889, 1893, 1897, 1900, 1912, 1916, 1919-1921, 1924-1931, 1938, 1939, 1958
 Health Informatics (Healthcare Information Systems) 1943
 Health Insurance Exchanges 1850, 1856
 Health Policy 1585, 1831, 1832, 1856-1858, 1873, 1874, 1882-1885
 Health System 1872, 1874, 1882, 1883, 1886, 1935, 1941, 1943
 Health Technology 1816, 1825, 1872, 1873, 1882, 1883, 1886, 1898, 1939
 Healthcare 28, 525, 528, 561, 574, 593-595, 745, 782, 787, 849, 868, 914, 917, 919, 921, 925, 1114, 1336-1338, 1342-1346, 1349-1352, 1366, 1554, 1562, 1599, 1686-1689, 1696, 1698, 1702, 1719, 1726, 1797, 1799, 1808, 1810, 1815, 1835, 1850-1862, 1867, 1868, 1872-1874, 1882-1886, 1894, 1899, 1909, 1919-1943
 Healthcare Monitoring 525, 917, 1943
 Healthcare System 868, 919, 1850-1856, 1867, 1919, 1920, 1923, 1941, 1942
 heat pipe 1272-1288
 Heat Transfer 310, 311, 1271-1272, 1274, 1278, 1281-1285, 1288, 1297, 1302-1303
 Heterogeneity 261, 265, 443, 577-579, 585, 639, 932, 1927, 1928
 heterostructures 1255, 1263, 1268
 Hexa-dimension Metric 411, 417-419, 422-427

hierarchical routing 260, 270
 High Dimension 576
 Highly Aged People 1944-1951
 H-O 662, 670, 674
 home area network 817, 1002, 1015-1016
 Home Area Network (HAN) 1002, 1016
 Honey Bee Algorithm 98
 Honeybee Algorithm 106
 Honeybee Mating Optimization 241, 244
 Hopping Windows 605, 618
 horizontal scaling 793, 799
 HTML5 1124-1125, 1128, 1517, 1525
 Human Bomb 355
 Human Germline Engineering 1966
 Human Rights 344, 398, 402, 404, 408, 448, 1481-1496, 1499, 1553, 1554, 1572, 1616, 1827, 1834, 1836, 1863
 human sensory assessment 1198-1199, 1205, 1208-1209
 Human-Agent-Robot Teamwork (HART) 27, 28, 41
 human-computer interface 147
 human-machine co-activity 28
 Hybrid Machine Translation 878, 882-886
 hybrid materials 1263-1266
 hybridity 1131, 1138, 1147
 Hydrodynamics 1289-1290, 1304, 1306-1307
 Hydrological Cycle 1358, 1361
 hyper-parameter 676, 678, 680, 683-684
 Hyponymy-Hypernymy 872
 Hypothesis 157, 160, 161, 212, 417, 562, 624-634, 639-641, 646-647, 652-654, 692-694, 700, 751, 753, 884, 1044, 1174, 1436, 1710, 1711, 1802, 1805, 1830, 1851, 1852, 1856, 1857

I

I2V communication 1053
 ICAO 658-662, 666-667, 671, 674
 ICT4D 1421, 1718
 IFR 668, 674
 Ijtihad 1860, 1862, 1870
 image acquisition 129, 133, 136, 137
 Image Analysis 22, 128, 130, 762, 1626, 1902, 1904, 1912
 Image Retrieval 87, 90, 95-97, 1781
 Incidental Endogeneity 577, 579, 585
 Inclination Angle 1271, 1279, 1282, 1287
 Individual Cloud 177, 178, 188, 189, 1944-1951
 Individualism 373, 1083, 1437, 1461
 inductive learning 751, 754-755, 767
 Industrie 4.0 1323, 1327, 1329, 1332, 1335

- INFARMED, I.P. 1816, 1817, 1820, 1826
 Infectious 1336-1350, 1399, 1726
 Inference 12, 23, 58, 68, 188, 333, 338, 500, 508, 509, 579, 581, 625, 627, 635, 636, 686, 691-693, 695, 698-699, 702, 712-713, 716-722, 767, 1242, 1249, 1471, 1477, 1480
 Inflectional Language 82, 86
 inflectional languages 73, 77, 79
 Information Age 431, 1451, 1453, 1675, 1765
 Information And Communication Technology 788, 818, 824, 968, 1067, 1106, 1418, 1421, 1422, 1660, 1701, 1715-1718, 1726, 1884, 1895
 Information And Communication Technology (ICT) 968, 1418, 1715, 1726
 Information and Communications Technology for Development (ICT4D) 1718
 Information and Information Literacy 1682
 Information Retrieval 87, 95, 96, 107-113, 116, 118, 122-125, 183, 683, 828, 844-846, 872, 881, 1049-1050, 1227, 1572, 1583, 1769, 1785, 1787, 1791, 1898
 Information Security 411, 412, 417, 420, 423-425, 468, 474, 577, 776, 1258, 1849, 1940, 1942
 Information Sharing 428-438, 448, 660, 674, 952, 1592, 1702, 1846
 Information System 229, 528, 542, 552, 561, 811, 1048, 1195, 1205, 1330, 1535, 1556, 1705, 1717, 1756, 1827-1831, 1836, 1838, 1884
 Information Systems Research 470, 1752, 1755, 1848
 Information Technology 9, 105, 177-180, 190, 214, 230, 242, 364, 392, 402, 422, 468, 525, 526, 656, 797, 812, 868, 908, 1055, 1082, 1103-1104, 1106, 1182-1183, 1228, 1418, 1420, 1452, 1457-1459, 1523-1525, 1564, 1594, 1679-1682, 1688, 1715, 1716, 1794, 1834, 1888, 1889, 1893, 1894, 1897-1901, 1941
 Informational Competence 1739, 1740, 1743
 Informational Competencies 1728, 1739-1742
 Infrastructure As A Service (IaaS) 463, 472, 1925
 Infrastructure-to-Vehicle (I2V) Communications 1052, 1064
 Innovation 586, 589, 932, 936, 1112, 1147, 1159, 1184, 1323-1326, 1333, 1359, 1361, 1400, 1421, 1424, 1427, 1430, 1433, 1434, 1440, 1449, 1455-1459, 1526-1536, 1562, 1563, 1595, 1602, 1697, 1724, 1764, 1776, 1777, 1780, 1793, 1873, 1930, 1950
 Input-Process-Output Model 1685
 Institutionalism 1852, 1854
 Insurance Market 1850, 1851, 1859
 intelligent ambient 139, 145, 147
 intelligent connectivity 1928, 1929
 Intelligent Driver Assistance Systems (IDAS) 11
 Intelligent Transportation Systems 6-10, 1051-1052, 1061-1062, 1064
 Intelligent Transportation Systems (ITS) 1051-1052, 1064
 Interactivity 350, 409, 547, 663, 1132, 1138, 1141, 1150, 1153, 1155, 1159, 1161, 1768
 internet addiction 1066, 1072, 1074, 1078, 1080-1081, 1083-1084
 Internet Archive 1462, 1782-1786, 1792
 Internet of Everything (IoE) 1932
 Internet of Medical Things (IoMT) 1932
 Internet Of Things (Iot) 28, 208, 215, 266, 537, 586, 589, 593, 594, 597, 601, 675, 887, 913, 922, 924-925, 935-937, 945, 1037, 1135, 1199, 1205, 1207, 1920, 1921, 1929, 1933, 1934, 1940, 1943, 1946
 Internet of Vehicles (IoV) 11
 Internet Service Provider 464, 482, 967, 1156
 internet television 1066, 1068-1072, 1074-1076, 1078, 1082, 1084
 Interoperability 140, 217-232, 598, 887-889, 924-926, 930-937, 1037, 1768, 1786, 1837
 Interpolation 290, 291, 310-313, 754
 Intersectionality 370, 389-397, 1101-1102, 1105, 1410, 1715
 Intrusion Detection Scheme 513, 538
 Inverse Distribution Function 173, 176
 inverted pendulum 42-45, 52, 53
 Investigator 335, 464-468, 1801-1804, 1809-1813, 1895
 Investment 266, 413, 419, 424, 1087, 1114, 1135, 1186, 1321, 1323, 1327, 1330-1332, 1415, 1416, 1504, 1505, 1527, 1532, 1533, 1570, 1571, 1604, 1662, 1665, 1668, 1671, 1675, 1690, 1694, 1714, 1733, 1734, 1894, 1923, 1957
 IoT interoperability 924, 930-931, 934-935, 937
 IoT Protocols 933, 937
 IRB 1802, 1803, 1807, 1810, 1811
 Islamic Medical Ethics 1860-1865, 1869, 1870
 island-style routing 260, 261, 267, 270
 Italian 107-110, 114-118, 121, 122, 126, 358, 376, 625, 1784
 Iv 10, 36, 107, 111, 280, 285, 348, 644, 677-678, 862, 1021, 1078, 1142, 1144, 1155, 1170, 1293, 1336, 1352, 1370, 1390, 1395, 1396, 1514, 1517, 1700, 1770-1774, 1781, 1789, 1799, 1818, 1864, 1879
- ## J
- Jennifer Doudna 1952
 Job Model 1555, 1557, 1560, 1574

Index

Journalism 1094, 1130-1150, 1152-1154, 1158-1163, 1196, 1461-1469, 1480, 1489, 1572, 1573, 1677
Journalist 1136-1137, 1140-1146, 1156-1157, 1461
JSON (JavaScript Object Notation) 499
Jump Ahead (Back) v Steps 176
Justification 1425, 1426, 1429-1433, 1805-1808, 1811, 1854

K

Kaldor-Hicks Criterion 1622
KDD Process 750-753, 766, 1923, 1924, 1928
KDM (Knowledge Discovery Metamodel) 890, 902
Key Competence 1740, 1744
Keywords 97, 107, 108, 111, 113, 118, 121-125, 617, 830, 1124, 1444, 1445, 1452, 1865
K-Means 241, 570, 575, 576, 759, 770, 830, 1914
Knowledge Base 183, 662, 697, 783, 1167, 1228, 1360, 1535, 1945
Knowledge Discovery 559, 573, 574, 597, 606, 615, 618, 681-684, 749-751, 764-765, 776, 786, 815-816, 818, 823-824, 845, 871, 890, 901-902, 1477-1480, 1923, 1924, 1929, 1930
Knowledge-Based Anomaly IDS 513
Kohonen 12, 1903, 1918

L

Landslide 1623-1626, 1629-1636, 1910, 1914
Landslide Hazard 1623, 1630, 1633-1636
Landslide Risk Assessment 1623, 1631-1636
Language Morphology 86
Law Enforcing Agencies 1481, 1484-1495
LCD 272-275, 280-285, 288, 289
LCD interfacing 274, 282, 288
LDA 112, 118-121, 774
Lemmatization 108, 109, 116, 121, 125, 862, 871
lender-borrower approach 801-802, 805, 810
Lexical Relationship 872
LFT 848-850, 852-853, 855
Liberalism 1453, 1460
Library 3.0 1769-1772, 1775-1781
Library Management 1765, 1775, 1780
Library Services 1767-1781
Light Detection and Ranging (LIDAR) 11
LINE Application 912
linear array design 953
Load Balancing 98-102, 105, 106, 224, 233, 237-244, 937, 1000, 1019-1020, 1022, 1037-1040, 1042, 1046-1050

load balancing cluster based web server 1038, 1040, 1042, 1046-1048, 1050
load testing 1040-1041, 1050
Local Government 1346, 1504, 1523, 1572, 1579, 1585-1588, 1592-1594, 1597-1602, 1605, 1606, 1610, 1611
Location-Based Services (LBSs) 1387, 1395, 1396
locking-committing conflict 809-810, 814
locking-locking conflict 805, 810, 814
locking-processing conflict 805, 810, 814
Logical Operations 271-275, 279, 282, 283, 286, 860
Logistic Regression 631, 633, 769-770, 847-848, 850-851, 853-854, 1634, 1832, 1903, 1910, 1914, 1916
Loneliness 360, 364, 373, 380, 389, 1081, 1089-1090, 1092, 1641
long short-term memory 96, 97
Look-Up Table (LUT) 270
LSI 112, 118-121, 272, 829
LSTM 89, 94-97, 586, 589

M

Machine Learning 12, 13, 22-26, 58, 74, 97, 122, 125-137, 140, 146, 210-216, 267, 443, 489, 559, 562, 569, 571, 574-578, 583, 584, 595, 596, 599, 604, 609, 614, 617, 658, 680-684, 749-752, 756, 762, 764-770, 777, 818, 845-848, 853-855, 878, 883, 946-947, 1311, 1312, 1472, 1473, 1657, 1834, 1896, 1897, 1901-1909, 1913, 1924, 1928, 1932, 1940
Machine Translation 12, 73-86, 762, 874, 876-886, 1207
Machine Translation Approaches 876, 884
Machine Translation Error 85, 86
Machine Translation Evaluation 86, 885-886
magnetic resonance imaging scans 1874-1881, 1886
Magnetic Resonance Imaging Scans (MRI) 1886
Mainlobe Beamwidth 966
Makerspaces 1762
malignant tumors 847, 855
Malta 1342, 1514, 1517, 1518, 1529
Malware 473-475, 485, 490-492, 929
Management component 550, 551, 554, 555, 558
Markup 207, 1037-1038, 1118-1119, 1123-1124, 1127-1129, 1525, 1781
Marxism 1453, 1460, 1646
Mass Communication 432, 1080, 1136, 1148, 1150, 1157, 1205, 1461, 1465-1468, 1715
Mass-Digitization 1782, 1785, 1791, 1792, 1795
Massive Digital Libraries (MDLs) 1782, 1783, 1793, 1795

- Mathematics 54-57, 60, 67, 70, 72, 174-176, 241, 242, 290, 310-314, 319, 328, 330, 750, 940-941, 961, 978, 1322, 1427, 1470, 1714, 1910
- maximum end-to-end delay 970-971, 1031
- MCA (Medical Care Act) 1859
- Means to Harm 1426, 1430
- Medicaid 1850-1859
- Medical Data 857, 1049, 1827, 1833, 1834, 1838, 1902, 1927
- Medical Device 1815-1819, 1822-1826, 1915
- Medical Network 1829, 1838
- Medical Networks 1827, 1829, 1833-1835
- MEMS 1948-1951
- meronymy 857, 859, 866, 872
- Messenger 375, 906, 908, 912, 1757, 1864, 1870
- Metacognitive Architecture CARINA 54-56, 65, 67, 71, 72
- meta-level 58, 60
- metamodeling 888, 892-893, 898, 900-902
- MHealth 1696, 1725, 1919, 1935-1938, 1941-1943
- Microbial 1368, 1376-1378
- Microphone Array 139-142, 148, 149, 152, 153
- Microsoft Excel Solver 1243, 1250
- Migration and Evacuation in Disaster 1622
- Migration: Transporting legacy systems to a newer platform (hardware or operating system) or language. For example 1230
- Mini Cloud 177, 185-188, 1946
- Minorization 344, 346, 355
- Mission Statement 1761, 1766
- Mitigation 413, 417, 423, 426, 526, 725, 1055, 1321, 1398, 1401-1404, 1407-1410, 1413, 1416, 1417, 1422, 1632, 1815, 1818
- mobile agent 500, 504-513, 517, 526
- Mobile Applications 230, 473, 476, 480, 489, 490, 879, 892, 897, 902, 926, 1067, 1070, 1516, 1524, 1595, 1689, 1725, 1840, 1841, 1846, 1849, 1939
- Mobile Communication 597, 601, 908-909, 1726, 1773, 1776
- Mobile Computing 215, 218, 230, 512, 525, 537, 798, 892, 1035, 1935
- mobile development 891-892
- mobile handheld devices 1384, 1395
- Mobile Health (mHealth) 1943
- Mobile Health and Fitness Applications 1839, 1845, 1846, 1849
- Mobile Phone 357, 489, 887, 901, 905, 908-910, 1199, 1419, 1590, 1702, 1708, 1720-1726, 1774, 1935
- Mobile Point of Care (MPOC) 1726
- mobile robots 42, 52, 58
- Mobile Technologies 398, 399, 887, 1199, 1205, 1208, 1515, 1595, 1648, 1719, 1750, 1773, 1781, 1941, 1943
- Mobile Technology 902, 1086, 1715, 1719-1723, 1726, 1781, 1935
- Mobile-edge computing 30, 41
- Mobilization 1100, 1103-1105, 1111, 1357, 1436-1440, 1443-1450, 1597, 1602, 1603, 1739, 1740
- Mobilization-Decision 1436, 1437, 1444-1450
- Mobilization-Decision Theory 1436, 1437, 1445-1450
- Model Checking 889, 1004-1006, 1013-1016
- Model Driven Architecture 900
- model transformation 890, 902-903
- Model-Driven Architecture (MDA) 888, 903
- Model-Driven Decision-Making and Control 1322
- Model-Driven Engineering 547, 901, 903, 1222
- modelization 548, 558
- moderation 1157-1158, 1160, 1162-1163
- Molecule 155, 730-731, 733-744, 1366, 1828
- Monitoring Applications 1726
- Moral Responsibility 1452
- Morally Permissible 1428-1431
- Morally Wrong 1424, 1425, 1428-1431, 1686
- Mosaicism 1959, 1966
- MQTT 926, 933-934, 936-937
- MR 1019, 1575, 1723, 1833-1836, 1890, 1891, 1897, 1951
- MRI 561, 587, 847-848, 855, 1872-1878, 1883, 1886, 1902, 1913, 1917
- MSB 278, 280, 289
- Multi Dipoles 177
- multicast routing 1018-1020, 1022-1024, 1026-1036
- multi-level communication 1463
- Multimedia 24, 115, 149, 206, 230, 242, 268, 473, 474, 477, 537, 538, 601, 618, 825, 892, 963, 980, 1018, 1063, 1134, 1140, 1144, 1147, 1149, 1182, 1694, 1702, 1774, 1776, 1781
- Multimodal learning 87, 96
- Multitasking 43, 49-56, 60, 1726
- multitenancy 463, 469, 472
- Municipality 1505, 1579, 1586-1591, 1672
- Mw 254, 1336, 1352
- MWC 39, 490, 513, 539, 1352
- MWM 1352
- MWMP 1352
- MXene 1259-1260, 1262-1263, 1266-1270
- MXenes 1254-1255, 1258-1260, 1262, 1266-1268
- Myspace 370, 384, 1098, 1106, 1462, 1585
- MySQL Databases 1385, 1386, 1397
- MySQL Workbench 1384-1386, 1395, 1397

N

Nano Materials 1272, 1288
 nanofluids 1271-1273, 1275, 1278, 1282-1288
 narbs 451-454, 459-461
 Narrative 451-455, 458-461, 465, 1082, 1132, 1134-1135, 1137, 1141, 1147-1148, 1605, 1639, 1704, 1896
 narrative map 452-454, 460
 Narrowband Internet of Things 913, 922-923
 Narrowband Internet of things (NBIoT) 923
 National Framework 1497, 1511
 National Innovation Systems 1535, 1536
 Natural Language Model 1179
 Natural Language Processing 12, 25, 70, 73, 74, 85-87, 107, 123, 658, 762, 767, 857, 868, 871-872, 883, 1472, 1473, 1477, 1896
 Navier-Stokes Equations 1291, 1307
 NBC 1087
 NBIoT deployment 916-917, 923
 neighborhood area network 817, 1002
 Neo4j 778-779, 781-783, 785, 787, 789, 794
 Network Cost Estimation 949
 network dimensioning 944, 949
 Network Journalism 1132, 1136, 1138
 Network Layer 26, 514, 515, 518, 522, 523, 530, 928, 980, 982-983, 1024
 Network Performance Management 979
 Network Performance Management System 979
 network science 939-943, 945-947, 949
 network structure 982, 1108, 1111, 1113, 1117
 Network Virtualization 935-936, 938
 Neural Machine Translation 76, 878-879, 882-886
 Neural Networks 12-18, 22-26, 51, 57, 71, 76, 88, 93, 97, 135, 140, 143-146, 151, 152, 291, 311-313, 582, 613, 625, 683, 755, 761-762, 765, 773, 820-821, 854, 866, 879, 886, 944, 1249-1250, 1284, 1383, 1473, 1901, 1902, 1908-1918
 News 160, 334, 344-347, 352-354, 406, 407, 424, 425, 430, 442, 451, 452, 460, 471, 561, 634, 863, 865, 882, 1069, 1079, 1082, 1087, 1090, 1092, 1094, 1099-1100, 1131, 1133-1136, 1138-1141, 1144, 1146-1148, 1152-1163, 1332, 1351, 1406, 1458, 1461-1480, 1487, 1492, 1510, 1520, 1523, 1550, 1572-1574, 1619, 1620, 1643, 1678, 1692, 1714, 1724, 1761, 1774, 1785, 1790, 1848, 1854, 1869, 1964, 1965
 NewSQL databases 796-798
 NFV 925, 932-938
 Ngram Viewer 1784, 1790-1792, 1795

Nigeria 344-347, 351-354, 399, 407, 967, 974, 976, 1231, 1336, 1340, 1342, 1348-1351, 1362, 1364, 1367, 1372-1374, 1420, 1699, 1700, 1703-1712, 1715-1717, 1725, 1726, 1767, 1779, 1780
 Node for Internet Access 973-974, 979
 Noise Accumulation 577, 579, 585
 non-computability 160, 161
 Non-Infectious 1336, 1337, 1343, 1346
 Nonlinear System 290, 291, 311, 314, 1249, 1284
 Normative Belief 373
 Norms 76, 363, 379, 398, 403, 404, 439, 905, 1071, 1111, 1131, 1156, 1158, 1399, 1400, 1406, 1407, 1410, 1414, 1416, 1421, 1452-1454, 1459, 1460, 1490, 1549, 1703, 1704, 1712, 1832
 Nosql 560, 574, 588, 667, 670, 674, 778, 782, 785, 791-792, 794-799
 NoSQL (Originally Referring to “Non-SQL” or “Non-Relational”) Database 674
 Nosql Database 560, 574, 667, 670, 796-797, 799
 Notified Bodies 1816, 1826
 N-Tuple Algebra 685-686, 698, 700
 nucleotide 730, 747
 Nudge 447, 1537-1539, 1546-1548
 numerical methods 728, 1291

O

OAE-QoS 200, 206
 OAuth 495-499
 Obamacare 1858, 1859
 obfuscation 473, 485, 491
 object-level 60, 66
 OFDM 950, 957-958, 960-963, 966
 OLS Regression 657
 on premise 208, 214, 216
 One-Way Error Model 657
 online behavior 335-343, 365, 397
 Online Content Providers 1091, 1095
 online harassment 385, 402
 Online Privacy Policy 1849
 Ontologies 57, 69, 71, 856, 858, 868-871, 1770, 1771, 1776, 1777
 Open Data 123, 598, 1142-1143, 1145, 1147-1151, 1311, 1516
 Open ID connect 497-499
 Open Source 794, 869, 879, 883, 893, 1129, 1134, 1143, 1185, 1187, 1195, 1386, 1397, 1455, 1516, 1668, 1792
 Operating Environment 903, 1560, 1562, 1574, 1609
 Operating Model 1353-1361
 OPS 982-983, 997, 1001

Optimal Control 962, 1309, 1310, 1315-1322
 Optimization 21, 23, 101, 105, 106, 158, 159, 162, 194, 215, 235-244, 252, 255, 531, 547, 577, 583, 601, 660, 663, 672, 684, 723, 782, 786, 789, 818, 824, 913, 922, 942, 945, 950-951, 953, 957, 959-966, 979, 998, 1020, 1023, 1026-1028, 1031, 1034-1035, 1248-1250, 1252, 1283, 1303, 1309, 1315, 1317, 1320-1322, 1349, 1376, 1477, 1787, 1908, 1923, 1941
 Optimize 42, 46, 67, 98, 214, 268, 634, 724, 821, 1290, 1327, 1329, 1551, 1554, 1563, 1574, 1927
 Optimized Sustainment Model (OSM) 1552, 1574
 Organ Transplantation 1861-1867, 1870, 1871
 Organizational Competence 1740, 1744
 Orthogonal Frequency-Division Multiplexing (OFDM) 966
 OTT 880, 884, 1086-1087, 1090, 1093, 1095, 1468
 Ozonation Treatment 1365, 1383

P

Panel Data 637-642, 644-645, 654-657, 1885, 1886
 PANS 674
 parallel clustering 559, 566-569, 572, 576
 Parallel Threads 216
 parametric uncertainty 701-702, 724, 726, 728-729
 parental mediation 361, 373, 378, 384
 Parental Mediation and Monitoring 373
 Parent-Child Relationships 392, 1074
 parenting style 362, 373, 378
 Parents 361, 362, 365-367, 373, 374, 378-382, 387, 392, 393, 403, 1077, 1085, 1554, 1639, 1643, 1750, 1833, 1853, 1957
 Participatory Democracy 1579-1582, 1586-1590, 1593-1595
 participatory journalism 1152, 1154, 1158-1161, 1163
 Particle Swarm Optimization 237-244, 951, 959
 patch antenna 950, 953-955, 960-963
 Patient Zero 1836
 PCA 630-633, 774, 1628, 1756
 Peacemaking 1464, 1466
 Pedagogy 1466, 1660, 1661
 peer attachment 360, 363, 367, 371, 373, 379, 382
 Peer Contagion 363, 373
 Peers 361-363, 366, 373, 374, 378, 379, 385, 388-392, 400, 1074, 1077, 1108, 1454, 1649, 1731
 Penetration 817, 820-821, 1067-1068, 1092, 1095, 1145, 1260, 1516, 1649, 1706, 1885
 Perception 57, 60, 69, 87, 129, 152, 335-339, 361, 367, 381, 382, 400, 407, 447, 450, 750, 1071, 1149-1150, 1156, 1173, 1192, 1197-1198, 1200-1201, 1203-1205, 1208, 1210-1211, 1411, 1416, 1439, 1445, 1446, 1449, 1489, 1514, 1516, 1634, 1645, 1646, 1734, 1739, 1753, 1754, 1918
 Perceptron 122, 761-762, 773, 1899, 1918
 Performance-Based Approach (PBA) 660, 674
 Period Length (of RNGs) 176
 Personal Data 421-425, 489, 493, 495, 763, 777, 1538, 1588, 1726, 1830, 1836-1846
 Personal Identifiers Information (PIIs) 1849
 Personalization 147, 148, 683, 1155, 1768
 personalized medicine 1933, 1940
 Petri nets 42, 43, 46-52, 502
 phosphorene 1263, 1266
 Physical System 161, 163, 1327
 Physiological consequences 1066, 1073, 1076
 pipe flow 1248-1250, 1253
 PIR Sensor 139, 153
 PK-Means 576
 Platform As A Service (Paas) 463, 472, 1925
 Platooning 6, 11, 1059
 Pluralistic democracy 1559
 Policy Model 1850, 1851, 1859
 Political Responsibility 1451, 1452
 polyglot persistence 796, 799
 polymers 1255-1256, 1262, 1264-1265
 Polynomial 167, 172, 290, 310, 313, 314, 691, 821, 1029, 1310, 1312, 1319, 1320
 polytope 733-736, 747
 Pooling Layer 17, 26, 1904
 Pop-Ups 1520, 1525
 POS (Part of Speech) Tagging 125
 Post-Editing 76, 83, 84, 875, 879-886
 Post-Truth 1469, 1470, 1476, 1477
 potsherd 129, 130, 133-137
 potsherds 127-137
 pottery 127-137
 Pottery Classification 128, 137
 Pottery Documentation 137
 Pottery Reassembly 127, 137
 pottery reconstruction 130, 134, 135
 Poverty 1359, 1403, 1405, 1417, 1421, 1422, 1482, 1500, 1503, 1555, 1597-1599, 1604, 1610, 1611, 1681, 1692, 1698, 1700, 1720, 1723, 1850, 1884
 Pragmatic Method 1212
 Precaritized Labour 1647

Index

Precision 80, 113, 114, 118-122, 168-170, 176, 266, 517, 525, 538, 539, 595, 848, 862, 864-866, 893, 945, 1052, 1227, 1275, 1328, 1335, 1769, 1772-1776, 1895, 1896, 1899, 1911, 1938
Precision Farming 1328, 1335
predicted task execution time 29, 33, 34, 37, 41
Principal Component Analysis 625, 630-636, 774, 1113, 1751, 1880, 1905
Privacy 141, 147, 148, 190, 204, 335, 341, 348, 361, 401-404, 409, 413-418, 421-449, 453, 454, 457, 458, 467-469, 473, 474, 486, 491, 528, 577, 581, 583, 763, 775-776, 908, 916, 936, 1133, 1202, 1204, 1394, 1396, 1643, 1704, 1776, 1782, 1783, 1789-1792, 1831, 1834, 1835, 1839-1849, 1895, 1899, 1920, 1927, 1936, 1940, 1942
Privacy Laws 423, 432, 433, 448, 1790
Private Sector 182, 1355, 1361, 1416, 1499, 1509, 1510, 1526, 1552-1555, 1563, 1575, 1596, 1609, 1663, 1691
process block 1170-1171, 1173, 1180
Process Commonality 1175, 1180
Process of Demographic Aging 1886
process variability 1167, 1175, 1180
Professional Competence 1740, 1744
profiler 334, 337-340
Prognostic 725, 1446, 1901, 1906, 1907, 1918
program understanding 1214, 1218-1219, 1228-1229
Proteomics 1888, 1895, 1897, 1900, 1909
Protocol 33, 49, 210, 215, 219, 399, 494-499, 502, 515, 517, 524-527, 530, 531, 677, 800-813, 843, 916, 925-928, 933, 935-937, 1002, 1004-1007, 1009, 1011, 1013-1017, 1020-1022, 1024, 1034, 1037-1038, 1124, 1403, 1773, 1797, 1798, 1801-1814, 1836, 1901, 1937, 1960
Protocol Deviations 1814
provictim attitudes 360, 364, 368, 373, 377, 382
P-Set 1749, 1756
Public Cloud 207-216, 467, 470, 617, 1926
Public Domain 457, 1784, 1785, 1788-1792, 1795
Public Finance 1497, 1504, 1510, 1597, 1604, 1607-1610
Public Health 399, 407, 429, 440, 539, 869, 1092, 1340, 1346-1351, 1366, 1375, 1412, 1504, 1538, 1619, 1812, 1823, 1829-1838, 1853, 1856-1858, 1884, 1935
public key 519-524, 527
Public Policy 374, 445, 447, 575, 1511, 1537-1542, 1545-1548, 1580, 1595, 1695, 1725, 1838, 1850-1857
Public Private Community Partnership 1355, 1361

Public Sector 182, 183, 1142-1143, 1331, 1497, 1506-1509, 1515, 1516, 1523, 1524, 1553, 1575, 1587, 1592, 1595, 1596, 1662, 1665-1668, 1686, 1689, 1714, 1859, 1885
Python 56, 116, 120, 124, 126, 162, 210, 214, 593, 656, 864, 868, 891, 1143, 1386, 1390, 1393, 1394, 1668, 1693, 1697

Q

Q-Method 1746-1755
Qos 191-194, 200, 205, 206, 926, 933, 1005, 1016, 1018-1021, 1023-1035, 1037-1040, 1048
Q-Sample 1747, 1756
Q-Sorting 1747, 1750, 1756
Qualitative Research 384, 1459, 1746, 1753, 1756
Quality Of Service 30, 192, 205, 206, 512, 926, 1018, 1026, 1028-1029, 1033-1034, 1441, 1928
Quantile Function (Inverse Distribution Function) 176
quantum computing 157, 597, 598, 601

R

R language 642, 655
RAB 1482, 1484, 1487, 1490, 1492, 1495
Radio Frequency Identification (RFID) 1775, 1920, 1932-1934
RAKE 110, 111, 118-121
RAM 209, 237, 243, 264, 270, 272, 289, 728, 983, 997, 1001, 1041, 1937
random effects 638, 640-641, 644, 649-651, 653-657
random number generator 164-169, 174, 175
Random Number Generators (RNGs) 176
ransomware 336, 413, 424, 425, 474, 491
Rater's Credibility 193, 206
Rational Resource 1850, 1851, 1857
RCW 1352
RDF 782, 784, 786, 1148-1149, 1225, 1227-1229, 1771, 1772, 1777
reactive jamming attack 528-537, 540
read receipt function 904, 906-907, 912
Read Receipt Status 908, 912
real-time commit protocols 801-802, 804-805, 814
Recall 80, 113, 114, 118-122, 679, 848, 862, 864-866, 973, 1227, 1404, 1428, 1470, 1615, 1617, 1772, 1776, 1840, 1896
Recommendation 147, 175, 193-195, 204, 213, 380, 457, 462, 655, 665, 675-684, 778, 782, 1124, 1128, 1446, 1541-1543, 1547, 1548, 1834, 1835, 1929
reconfigurable 254-257, 262, 265-270, 273, 288
reconfigurable architecture 268

- Reconfigurable Computing (RC) 270
 reconfigurable hardware 270, 288
 recruitment 1491, 1806
 Recurrent Neural Networks 76, 93, 97, 762
 reengineering 890, 892-893, 901, 1165, 1170, 1174, 1214-1218, 1222, 1225, 1227-1230
 Refactoring: Changing the source code and preserving the behavior (in the object-oriented programming). 1230
 Reflection 143, 337, 849, 1056, 1134, 1138, 1256-1257, 1260, 1262-1263, 1265, 1426, 1452, 1735, 1742
 Regional disparities 1872-1874, 1882-1886
 Regional Disparities in Health (Health Inequalities) 1886
 Registration 133, 195, 672-673, 1021-1022, 1157, 1163, 1505, 1799, 1812, 1895
 Regression 310, 579, 584, 625, 631, 633, 637-642, 644-645, 656-657, 663, 678, 682-683, 754-755, 759, 769-772, 774, 777, 818, 820-821, 826, 847-848, 850-851, 853-854, 1044-1045, 1273, 1473, 1633-1636, 1832, 1886, 1903, 1909-1917
 Regression Algorithm 770, 777
 Relational Database 778, 782, 786-787, 789, 792-793, 796, 799, 1228, 1386, 1397
 relational model 791-794, 796-797, 799, 1218
 Reliability And Security 512, 723
 Remotely Piloted Aircraft (RPA)/Remotely Piloted Aircraft System (RPAS)/Unmanned Aerial Vehicle (UAV) 674
 Remotely Piloted Aircraft System 674
 Renewable Energies 1323-1326, 1329-1332, 1335
 Renewable Resources 1309, 1310, 1316, 1669, 1670
 Replication 90, 91, 242, 606, 614, 743-744, 795, 797, 799, 813, 1752, 1960
 Reply 480, 515-518, 523, 551, 904, 906-907, 909, 912, 1019-1020, 1022
 Representative Democracy 1581-1588, 1593, 1595, 1599
 Reputation 190-206, 401, 577, 1472, 1591, 1692, 1760
 reputation prediction 191, 192, 201-205
 Requirements Engineering Process 1165, 1168, 1174, 1176-1178, 1180
 Research Method 1212, 1746, 1752, 1755
 ResNet 20-22
 ResNeXt 21
 Resource Mobilization 1100, 1103-1105, 1438, 1446, 1449, 1602
 Resource Mobilization Theory 1100, 1103, 1105, 1446
 Response Time 56, 99, 192, 233, 240, 244, 811, 906, 967, 979, 1039-1041, 1043-1045, 1047, 1927
 REST 3-6, 58, 217-224, 228-231, 273, 316, 421, 435, 454, 495, 550, 689, 783, 789, 795-796, 799, 926, 1293, 1384, 1413, 1554, 1595, 1599, 1617, 1848, 1944
 Restricted Boltzmann Machines 87, 90
 Restructuring: The transformation from one structure to another at the same relative level of abstraction 1230
 Reverse Engineering 890-893, 896-897, 900, 902, 1214, 1216-1217, 1225, 1227-1230, 1908
 Reynolds-Averaged Navier-Stokes Equations 1307
 Right To Privacy 409, 432, 447, 458, 1834, 1835, 1842
 Risk Management 1612, 1613, 1623, 1716, 1816, 1923, 1939
 robot journalism 1130, 1133-1134, 1137-1138
 Robotics 7, 30, 40, 52, 57, 130, 214, 215, 525, 658, 749-750, 752, 763-767, 1327, 1933
 ROM 264, 272, 289
 Room Acoustic 153
 Root Mean Square and Mean Absolute Error 206
 Round-Trip Time 968, 970, 972, 979
 Round-Trip Time (RTT) 968, 970, 979
 Routing Protocols 517, 825, 1002, 1004-1007, 1013-1015, 1033
 RPS 667-668, 670, 674, 1048
 RS 282, 787, 959, 1211, 1348, 1623, 1624, 1631-1633, 1636
 Rule-Based Machine Translation 876, 886
- ## S
- Safety Management System (SMS) 674
 SARPs 674
 Satisfaction 33, 99, 227, 364, 380, 401, 405, 428, 586, 589, 1074, 1094, 1112, 1516, 1552, 1553, 1564, 1571, 1589, 1736, 1741, 1830, 1872, 1889, 1923
 Scalability 92, 221, 260, 463, 464, 481, 490, 559, 563, 566, 569-572, 577, 578, 581-585, 675-676, 678, 680, 798, 822, 928, 1021, 1039, 1049-1050, 1053, 1768, 1782, 1925, 1927
 Scaling 211, 212, 287, 568, 570, 781, 791, 793, 798-799, 921, 940, 946, 1442, 1554
 Scenario 2, 107, 147, 193, 194, 221, 263, 501, 515, 531, 775, 802, 874, 887, 1018, 1024, 1026, 1033, 1038, 1058-1060, 1072, 1110, 1131, 1172, 1178, 1180, 1184, 1192, 1207, 1355, 1357, 1412, 1635, 1832-1835

Index

- Scenarios 31, 75, 140, 231, 502, 515, 581, 599, 680, 913-914, 917, 919, 932, 1002, 1032, 1052-1053, 1055-1056, 1060, 1166, 1168, 1170-1172, 1177-1179, 1182-1189, 1192, 1194, 1454, 1527, 1532, 1618, 1619, 1693, 1742, 1831, 1834
- School 190, 290, 356-397, 401, 402, 405-407, 428, 439, 440, 537, 556, 586, 631, 746, 885, 904, 1074, 1077, 1079-1081, 1114-1115, 1141, 1148, 1189, 1271, 1377, 1409, 1411, 1419, 1457, 1471, 1547, 1554, 1560, 1562, 1572, 1623, 1658, 1666, 1668, 1690, 1694, 1699, 1703, 1713, 1742, 1758, 1759, 1764, 1797, 1960
- Schools 346, 356, 361-370, 379, 382-384, 387-396, 402, 403, 406, 1145, 1186, 1409, 1415, 1451, 1453, 1471, 1554, 1603, 1606, 1611, 1663, 1665, 1699, 1707, 1715, 1717, 1756
- Screening 855, 1070, 1079, 1266, 1853, 1894
- SDN 925, 932-935, 937-938
- Second-Screen Experience 1090, 1095
- Section Autonomy 1611
- Security Methodology 527
- Sedentary Behavior (SB) 1088, 1095
- Selected Factors 1363, 1371
- selfie journalism 1130, 1132, 1137-1138
- Semantic Knowledge Representation 54-57, 65, 66, 72
- semantic memory 54-58, 61, 64-67, 72
- Semantic Web 231, 786, 1049, 1145, 1148, 1157-1158, 1214, 1225, 1227-1228, 1479, 1771-1781
- Semi-Supervised Learning Algorithms 762, 770, 777
- SENet 22
- Senior Citizens 1085-1087, 1691, 1716
- Sensitivity 45, 851, 853, 855, 1060, 1199, 1416, 1905, 1918
- Sensor Network 500, 510-514, 524-527, 531, 538, 812, 820-823, 825, 827, 920, 1913, 1932, 1935, 1941, 1943
- Sensor Node 501, 1943
- Server 28-31, 41, 51, 99, 106, 167, 194, 207-209, 217-228, 232, 338, 463, 465, 472, 473, 479-484, 495-498, 553, 560, 605, 609, 610, 785, 799, 879, 916, 927, 1037-1043, 1046-1050, 1384-1388, 1391, 1392, 1396, 1774, 1775, 1781, 1844, 1891, 1935
- Service Delivery 1038-1039, 1047, 1507, 1553, 1571, 1597, 1601, 1602, 1605, 1607, 1611, 1689, 1714
- Service-Level Agreement (SLA). 472
- Service-Oriented Computing 1050
- SGML 1118-1120, 1124, 1129
- Shahādah 1860, 1870
- Shape Grammar 1193, 1195
- sharding 795, 797, 799
- shared memory 48, 49, 53
- Shari'ah 1860, 1862, 1867, 1870
- Sharp Power 1471, 1474, 1480
- shielding effectiveness 1254, 1256, 1260, 1262-1267, 1270
- shielding effectiveness (SE) 1254, 1256, 1260, 1262-1264
- Short Message Service (SMS) 1726
- Sidelobe Level (SLL) 966
- Signal Strength 516, 519, 520, 524, 527-535, 1000, 1027
- Signs 6, 11, 401, 431, 595, 1102, 1127, 1197, 1200, 1209, 1688, 1889, 1890, 1896, 1897, 1900, 1919, 1928
- simplex 734, 736, 747, 1184
- Simulation 7, 50, 68, 70, 155-170, 173-176, 187, 188, 229, 247, 252, 310-313, 500, 508, 529, 532, 533, 547, 575, 634, 659, 701, 718-719, 721, 723, 726-727, 801, 807, 960, 974, 1005, 1014, 1188, 1208, 1210, 1212, 1248, 1250, 1285, 1304-1307, 1322, 1475, 1613, 1618, 1621, 1627, 1635, 1640, 1641, 1836, 1941
- sinkhole attack 514-520, 523-527
- Situational Factor 1180
- Situational Method Engineering 1165, 1176-1180
- sixdegrees.com 1462
- skip commercials 1072, 1075-1076, 1078, 1084
- Slice 258, 259, 1204, 1221
- sliding block 251-254
- Sliding Windows 604-607, 614, 615, 618, 676-679, 681
- slow journalism 1134-1135, 1137-1139
- Smart 3, 4, 8, 28, 99, 128, 138-153, 187, 188, 266, 421, 474, 500, 502, 511-514, 531, 536-539, 586-589, 592-600, 659, 670, 674, 698-699, 811, 815-826, 892, 902, 908, 914, 917, 919, 921-923, 925-926, 933-934, 942, 945, 1002-1005, 1007, 1009, 1013-1016, 1040, 1049, 1064, 1068, 1075, 1087, 1091-1092, 1130-1131, 1139, 1145, 1158, 1187, 1194, 1196, 1198, 1207-1209, 1211-1212, 1285, 1311, 1321, 1326-1331, 1424, 1467, 1472, 1531, 1534, 1586-1589, 1593-1595, 1640, 1687, 1719, 1722, 1726, 1768, 1770, 1774, 1836, 1921, 1922, 1933-1942, 1946
- Smart Cloud 187, 188
- Smart Environment 128, 142, 513, 1207, 1933
- Smart Grid 511, 531, 537, 587, 592, 599, 815-826, 923, 945, 1002-1003, 1005, 1013-1016, 1040, 1049, 1939
- smart grids 826, 1002, 1004-1005, 1007, 1013-1016
- Smart Home 138-141, 144-148, 151-153, 500, 536, 594, 917, 1014
- smart technology 1130, 1139
- SNS 185, 1083, 1107, 1109-1110, 1112

- SOA 181, 208, 217-224, 229, 231, 986-989, 993, 998, 1037, 1229, 1946
- Social Capital 395, 430, 444, 1106-1107, 1109-1117, 1420, 1677, 1690, 1698, 1835, 1836
- Social Exclusion 357, 373, 375, 1703
- Social Media 333-335, 351, 369, 374, 387-390, 393, 397, 400-404, 409, 430-438, 448, 453, 459, 467, 579, 605, 617, 637, 909, 912, 1067, 1070, 1082, 1090-1091, 1096-1105, 1115-1116, 1131-1132, 1134-1136, 1138-1139, 1152, 1155, 1157-1158, 1160, 1162, 1461-1480, 1550, 1573, 1579, 1585, 1586, 1595, 1613, 1645, 1651, 1661, 1676-1679, 1683, 1686, 1687, 1691, 1693, 1700, 1719, 1762-1770, 1830, 1831, 1841
- Social Media Tools 433, 448
- Social Movement 353, 1096-1097, 1099-1102, 1104-1105, 1436-1439, 1446-1450, 1479
- Social Movements 1096-1097, 1099-1104, 1437, 1438, 1447-1450, 1475
- Social Network 348, 363, 437, 441, 456, 459, 577, 762, 941, 1106-1109, 1111, 1114-1117, 1142, 1421, 1462, 1466, 1473, 1478, 1827-1838
- Social Networking 341, 356, 357, 366, 367, 371, 375, 381, 384, 385, 398, 399, 403, 407, 444, 455, 459, 476, 477, 559, 568, 908, 1098-1099, 1105-1107, 1109-1111, 1113, 1115, 1117, 1146, 1152, 1156, 1384, 1458, 1466, 1686, 1697, 1759, 1772, 1779
- Social Networks 22, 147, 399, 443, 457, 459, 486, 675, 778, 782, 939-941, 944, 947, 1090, 1104, 1106-1107, 1109-1110, 1114-1116, 1155, 1213, 1394, 1395, 1415, 1462, 1466, 1469, 1477, 1519, 1591, 1677, 1716, 1725, 1827-1838, 1935
- Social Responsibility 419, 445, 447, 1452, 1454, 1457-1459, 1465, 1466, 1505, 1602
- Social Stratification 581, 1110, 1672, 1675, 1682, 1700
- social theory 1448, 1449, 1467, 1621, 1660
- Soft Distributed Real-Time Transaction(Soft DRTT) 814
- Softmax Function 15, 26
- Software Application 1649, 1653-1655, 1661, 1775
- Software As A Service (Saas) 463, 472, 1037, 1925
- Software Engineering 52, 230, 231, 240, 243, 495, 544, 555, 556, 813, 887, 889, 900-901, 903, 1033, 1165, 1176-1178, 1214-1215, 1223-1225, 1227-1230
- Software Literacy 1648-1656, 1659, 1661
- Software Maintenance 487, 1214, 1230
- Software Maintenance: Modifying a software system after delivery to correct its faults 1230
- Software migration 889, 903, 1230
- software modernization 889-891
- Software Platform 1142, 1661
- Software Requirements 1166, 1168, 1178-1180, 1847
- Solid Waste Management 1336, 1337, 1347, 1350-1352
- space syntax 1198-1201, 1203, 1208-1210
- spatial analysis 1183, 1196-1197, 1199-1200, 1212
- Specificity 23, 595, 851, 853, 855, 1415, 1586, 1905, 1918
- speech recognition error 84-86
- Speech Recognition Vocabulary 86
- speech signal 140-144, 147
- Speed of Reply 909, 912
- Sponsor 1804, 1806, 1809, 1811, 1814
- spurious correlation 577, 579, 585
- Spyware 473-475, 492
- SQL 476, 484, 553, 554, 605, 606, 618, 785, 790-791, 794, 796, 798-799, 892, 901, 1143, 1218-1219, 1386, 1397, 1891
- SQL language 794, 799
- SQLite Databases 1397
- Stages 60, 133, 275, 348, 501, 502, 530, 547, 549, 587, 592, 593, 606, 627, 660, 662, 678, 753, 768, 818, 844, 848, 853-854, 877, 897-898, 928, 930, 936, 943, 1072, 1076, 1097, 1101, 1135, 1140-1142, 1144-1146, 1153, 1155, 1158, 1168, 1174, 1189, 1197, 1439, 1443, 1515, 1539, 1546, 1554, 1587, 1610, 1752, 1889, 1954
- Stakeholder 418, 1116, 1310, 1357, 1358, 1361, 1378, 1587, 1622, 1900
- stalking 398, 401, 409
- State Mufti 1860-1868
- State of Nature 1497-1502, 1511, 1639
- Statistical Inference 579, 625, 635, 636, 702, 718-722
- Statistical Methods 109, 155, 577, 726, 875, 1247, 1807, 1905
- Statistical Modeling 140, 146, 940, 948-949, 1916
- statistical tolerance limit with expected $(1-\alpha)$ -confidence 702, 704, 714, 716-721, 729
- statistical tolerance limits 702-704, 714, 722-726
- statistical γ -content tolerance limit with expected $(1-\alpha)$ -confidence 702-703, 709, 711-713, 718, 722, 729
- Stemming 80, 108, 109, 118, 124, 125, 218-220, 1154
- Stickers 912
- Streaming 141, 147, 207, 209, 602-610, 613-618, 679, 683, 1067-1068, 1075-1077, 1081, 1084, 1086-1087, 1090-1091, 1093-1095, 1759
- Streaming Data System 602, 618
- Streaming Mining 618
- structural equation models 625, 632-636
- Study Protocol 1802, 1805, 1814
- Subjectivity 339, 1421, 1731, 1746, 1748, 1752-1756
- subordinate component 551, 558

Sub-Saharan Africa 1400, 1414, 1418, 1450, 1597, 1703, 1719-1726
 suicide 344-347, 351-355, 364, 366, 383, 389, 407, 461
 Supervised Learning 87, 130, 146, 762, 769, 777, 1902-1904, 1918
 supervised learning algorithms 769, 777
 supervised training 15, 26
 Surveillance Capitalism 1782, 1789, 1790, 1794, 1795
 Sustainability 179, 419, 424-427, 582, 631, 656, 914, 921, 942, 1052, 1150, 1326, 1334, 1349, 1350, 1359-1361, 1369, 1420, 1421, 1572, 1589, 1606, 1608, 1872, 1882, 1883, 1886
 Sustainability of the Health System 1872, 1883, 1886
 SW 258, 260, 1027, 1352
 Sweden 358, 369, 383, 384, 399, 406, 441, 525, 886, 1148, 1342, 1366, 1440, 1514, 1517-1520, 1594, 1930
 Switched Local Area Network 978-979
 Symptom 412, 1890, 1900
 Synchronization 38, 46-49, 52, 53, 164, 168, 251, 252, 487, 571, 795, 1007, 1009-1010
 System (in Management) 331
 System Reliability 237, 240, 244, 538
 System-Wide Information Sharing and Management (SWIM) 660, 674

T

Tactile Internet 27-32, 38-41, 1928, 1932
 task migration 29-41
 task offloading 29, 31, 41
 tasks scheduling 233-241, 244
 Teachers 361-365, 368-370, 379, 382, 384, 391-394, 494, 661, 1073, 1077, 1458, 1602, 1660, 1687, 1717, 1754, 1786
 Technological Alienation 449, 460
 Technological Paradigm 1683, 1687
 Technological Policy 1526-1528, 1532, 1533, 1536
 Technology Use 356, 360, 361, 365, 390, 1094, 1418, 1680, 1682, 1704, 1715, 1749
 Telecommunication Market 1537, 1539
 Telehealth 1919, 1936, 1941
 Telemedicine 919, 1048, 1919, 1934-1936, 1941, 1943
 Television 214, 444, 445, 924-925, 966, 1007, 1066-1072, 1074-1076, 1078-1095, 1130, 1153, 1159, 1409, 1422, 1461, 1462, 1465, 1638, 1644, 1726, 1785
 television consumption 1066, 1069, 1072, 1074-1075, 1078, 1085-1086, 1088, 1091, 1095
 Terrorism 4, 333, 344-356, 374, 387, 398, 1481, 1644, 1646, 1698

Test Case 1041-1043, 1050, 1670
 TeX 1118, 1125-1126
 Text Mining 107, 123, 124, 828-829, 846, 856, 858, 868, 870-872, 1784, 1794
 text-based communication 904, 906-907, 909
 text-based query 87
 TextRank 110, 111, 118-123
 Tf-Idf 92, 110-112, 118-121, 125, 830
 Thana-Capitalism 1643, 1647
 The End of the World 1639, 1643-1646
 Theorem CAP 799
 Theory of Apocalypse 1638-1641, 1644, 1647
 thermal resistance 1273, 1276-1278, 1280, 1282-1283, 1288
 thermosyphon 1272-1273, 1277, 1282-1285, 1288
 Threat 4, 354, 381, 415, 457, 474-478, 483, 490, 492, 530, 540, 605, 1330, 1336, 1366, 1428, 1452, 1487, 1639, 1688, 1721, 1784, 1960
 Threats 357, 375, 400, 411, 415, 473-477, 481, 484, 485, 488-491, 501, 508, 510, 604, 1157-1158, 1328-1332, 1428, 1456, 1457, 1489, 1571, 1606, 1624, 1629, 1641, 1691, 1736, 1742, 1834
 Throughput 8, 30, 99, 210, 233, 250-255, 583, 593, 603, 607-611, 804, 967-968, 978-979, 998, 1018-1019, 1026, 1029-1032, 1039-1041, 1043-1045, 1047, 1054
 tie strength 1108, 1111
 Time Series 88, 637-639, 820, 826
 timed automata 1007, 1017
 time-shifting 1087, 1095
 TP 850, 1043, 1352
 traceback unit 248
 Trace-Back Unit (TBU) 256
 traditional face-to-face bullying 356, 357, 360, 364, 373-381
 Trajectory 43, 49-53, 381, 660, 674, 1291, 1293, 1297, 1353, 1359, 1961
 Trajectory Tracking 51, 53
 Transition Metal Dichalcogenide 1270
 transition metal dichalcogenides 1255, 1258-1259, 1266
 translation memory 879, 886
 Transmission 29-33, 38, 209-214, 245, 254, 255, 261, 373, 431, 477, 480, 500, 501, 505-510, 516-524, 527-536, 607, 816, 818, 820, 824, 914, 925, 927, 942, 944-945, 970, 973, 976, 981-983, 988, 1000, 1002, 1009, 1011, 1013, 1023, 1025-1026, 1031, 1051-1054, 1056, 1060, 1064, 1122, 1257, 1271, 1275, 1368, 1369, 1400, 1701, 1937, 1938
 Transparency, 1150, 1498, 1507, 1508, 1585, 1586, 1701

trellis decoding 245
 trellis diagram 246-251, 256
 Trials 957, 1617, 1797-1800, 1811-1813, 1816, 1823, 1894, 1952, 1955, 1959, 1960, 1963
 Triangular Fuzzy Number (TFN) 331
 Triangular Fuzzy Numbers 316
 Tripartite Link 427
 Trojan Horses 474, 492
 Trolling 398, 409
 Troubled Economy 1670
 Trust 190-195, 203-206, 415-418, 427, 432, 453, 493, 499-502, 505-508, 511, 514, 517-519, 522, 523, 538, 594, 775, 1106, 1111-1112, 1114, 1185, 1355, 1359, 1452, 1456, 1470, 1472, 1476, 1477, 1492, 1497, 1502, 1508, 1510, 1525, 1579, 1580, 1589, 1591, 1598, 1611, 1643, 1665, 1726, 1777, 1787, 1792, 1895, 1927, 1939
 TTC 669, 674
 Tumbling Windows 605, 618
 tumor markers 847-850, 853-855
 Tuning 95, 207-211, 216, 676, 680, 683, 986, 1262, 1561
 TV consumption 1088
 TWC 963, 986, 988, 998
 Twitter 333, 334, 399, 433, 452, 495, 560, 593, 1098-1102, 1104-1106, 1109-1110, 1112-1113, 1115, 1142, 1152, 1462, 1465, 1468-1472, 1477, 1480, 1579, 1585, 1638, 1651, 1762, 1768, 1772, 1782, 1829, 1837, 1844
 two-dimensional material 1259, 1262, 1270
 two-parameter Weibull distribution 701, 707-708, 711, 713, 715, 717, 720, 727, 729
 Two-Way Error Model 657
 type II censored data 702, 726, 728-729

U

Ubiquitous Computing 419, 537, 600, 1213, 1226, 1229, 1929, 1932
 UDP 210, 215, 916, 926, 928, 933-934, 936
 Ultrasonography 847, 855
 unification of presentation forms and processing procedures 685
 Unique Cloud 177, 188, 189
 un-oriented documents-words incidence matrix 832, 840, 843, 846
 Unread Receipt Status 912
 unsupervised keyword extraction 107
 Unsupervised Learning 24, 74, 130, 563, 758, 761-762, 769-770, 773, 777, 828, 1473, 1902, 1903, 1918
 unsupervised learning algorithms 769-770, 777
 unsupervised training 15, 16, 26

UPPAAL model checker 1004-1007, 1011, 1013, 1017
 Urban Emotion 1206, 1212
 urban form 1196-1197, 1212
 urban sensing 1196, 1199, 1205, 1208, 1213
 Urban Survey 1181, 1183-1184, 1189, 1192-1193
 Use of ICTs 1414, 1417, 1419, 1452, 1455, 1594, 1670, 1699-1706, 1709-1718
 User-Generated Content (UGC) 1152, 1163
 Utilitarianism 1453, 1460
 Utility 9, 176, 381, 394, 419, 426-448, 562, 582, 624, 628, 817, 854, 917, 969, 974, 977, 1030, 1059, 1069, 1114, 1309-1322, 1460, 1617, 1620, 1734, 1735, 1829, 1884
 Utility Independence 1311, 1322
 Utility Theory 428-433, 436-438, 441, 448, 1309, 1310, 1321, 1322
 Utilization 98, 99, 233, 262, 264, 413, 419, 424, 524, 562, 588, 816, 820-821, 918-919, 942, 983, 1002, 1029, 1038-1041, 1043-1045, 1047, 1135, 1250, 1290, 1304, 1309, 1320, 1379, 1410, 1538, 1542, 1682, 1752, 1805, 1808, 1873, 1885, 1893, 1921, 1923

V

V2I communication 1058
 V2V communication 1053, 1056
 VAD 140, 143, 144, 148, 153
 Value Driven 1320
 Value-Driven Design 1321, 1322
 variance reduction 168, 176
 variation point 1170, 1172-1173, 1175, 1180
 Varimax 1752, 1756
 Vehicle to Infrastructure (V2I) 11
 Vehicle-to-Infrastructure (V2I) Communications 1064
 Vehicle-to-Vehicle (V2V) Communication 11
 Vehicle-to-Vehicle (V2V) Communications 1064
 Verification 167, 466, 503, 547, 624, 627, 632, 694, 888, 900, 1002, 1004-1007, 1011, 1013-1016, 1147, 1157-1158, 1160, 1171, 1186-1187, 1470-1475, 1506, 1539, 1612, 1771, 1790, 1807, 1927
 vertical scaling 791, 793, 799
 VFR 674
 VGG 19, 21
 VHDL 273, 282, 287-289
 vicious circle 180, 411, 412, 427
 victimology 335-340, 343
 Video Streaming 207, 1075-1077, 1084
 video streaming portals 1075-1077, 1084
 Virtual or virtualized 472

Index

Virtual Reality 159, 162, 910, 1130-1131, 1137, 1139, 1183, 1194, 1205, 1207, 1451-1456, 1460, 1657, 1762, 1774, 1944, 1949, 1951
Virtual Reality Journalism 1139
Virtual Reference 1770, 1774, 1780, 1781
Virtual Reference Services 1770, 1774, 1781
Virtual Servers 99, 106, 464
Virtual World 187, 188, 334, 399, 909, 1196, 1198-1199, 1201, 1454, 1774
Virtualization 98, 190, 463, 597, 925, 929-938, 1768, 1776
Virtue Ethics 421, 426, 1453, 1460
Viruses 474, 492, 730, 744, 1938, 1954, 1964
Visible Light Communication (VLC) 1051, 1064
Viterbi algorithm 246, 247, 251-256
Viterbi decoder 146, 245, 247, 251-256
Viterbi Decoder (VD) 256
VLC receiver 1053-1054, 1059, 1064
VLC transmitter 1053-1054, 1059, 1064
VLOS 674
VLSI 254, 255, 267, 271, 272, 288, 289
Voice Activity Detection 140, 143-146, 149-152
VR 159, 162, 1130-1133, 1657, 1944-1951
Vulnerabilities 473, 477, 478, 484, 485, 490, 499, 536, 1472, 1475, 1522, 1615, 1626, 1927
Vulnerability 347, 354, 361, 365, 415, 473, 474, 479, 484, 486, 491, 492, 1080, 1360, 1399, 1401, 1405, 1408-1412, 1415, 1420, 1470, 1520, 1624-1629, 1632-1635
Vulnerability 1636

W

Waggle Dance 106
WAI 1515, 1524, 1525
Waste Generation 1342-1345, 1348-1351
Water Management 1355-1361, 1413
Water Security 1355, 1361
wayfinding 1197, 1199-1200, 1203, 1208-1211, 1213
Way-Finding 1197
WCAG 1515, 1516, 1522-1525
WDM 32, 984, 998-1001
Weak AI 767
Wealth Distribution 1692, 1693
Weapons Research 1424-1426, 1429-1435
Web 3.0 Technologies 1767-1769, 1774-1778
Web Authorization 493
Web Content Accessibility Guidelines (WCAG) 1515, 1523, 1525
Web Service 204, 205, 471, 539, 1037-1038, 1048-1050, 1392, 1393, 1767, 1840

Web Technologies 1039, 1145, 1158, 1225, 1514, 1767, 1774, 1777, 1781
wheeled inverted pendulum mobile robot 43, 45, 52, 53
whitenization 319, 320, 327, 331
Whitenization of a gn 331
Wide Area Network 31, 817, 923, 933, 1002, 1934
Wi-Fi 139, 477-481, 887, 966, 1067, 1440, 1932, 1933
Windows Join 607, 609, 618
Wireless Communications 39, 215, 231, 255, 256, 271, 490, 511-513, 536, 539, 599, 825, 874, 887, 904, 913, 921, 924, 939, 947, 950, 953, 960, 963, 967, 978, 980, 1002, 1018, 1035, 1037, 1051, 1055
Wireless Local Area Network (WLAN) 966, 1005, 1932
Wireless Mesh Network 1004, 1018, 1021, 1025, 1033-1035
Wireless Network 11, 481, 491, 530, 539, 540, 1005, 1014, 1440
Wireless Sensor Network 500, 511-514, 524-527, 531, 538, 812, 820-823, 825, 827, 1913, 1932
Wireless Sensor Network (WSN) 513, 527, 1932
Wireless Sensor Networks 208, 215, 239, 511-513, 524-527, 531, 537, 538, 675, 816, 820-822, 824-825, 915, 961, 964, 1004, 1014, 1212, 1412, 1943
Wireless Sensor Networks (WSN) 208, 1943
word sense disambiguation 83
Workflow 27, 39, 240, 243, 604, 1136, 1142, 1157, 1218, 1388-1391, 1889-1894, 1897, 1901, 1905
workplace cyberbullying 387, 389, 394-397, 400, 405
World Bank 287, 1325, 1409, 1414, 1415, 1418-1422, 1498, 1503, 1514, 1524, 1538, 1547, 1663, 1664, 1720, 1724, 1816, 1825
Worms 474, 492
WS Scalability 1050
WS Stability 1050

X

XML 187, 207, 210-214, 219, 221, 230, 553, 894, 1037-1038, 1118-1120, 1123-1124, 1128-1129, 1217-1218, 1391, 1772
XOR gates 264, 275-277, 280, 289

Z

ZigBee routing protocol 1005, 1007, 1017
ZigBee-based routing protocol 1002, 1004, 1006-1007, 1009, 1011, 1013, 1015
Zimbabwe 1421, 1497, 1499, 1503-1511, 1597-1611
Z-number 291, 296-314